Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.9, No.7, 2018
www.iiste.org

Real-Time Stock Trend Prediction via Sentiment Analysis of News Article

Sanmoy Paul Shashank Vishnoi
Department of Data Science (Business Analytics), NMIMS University, Mumbai sanmoy

Abstract
The stock market is volatile and volatility occurs in clusters, price fluctuations based on sentiment and news reports are common. A trader uses a wide variety of publicly available information to forecast the marketing decision. This paper proposes an advice to traders for stock trading using sentimental analysis of publically available news reports. It is based on a hypothesis, that news articles have an impact on the stock market, with this hypothesis we study the relationship between news and stock trend and also proved that negative news has a persistent effect on the stock market. In order to prove this assumption semi-supervised learning technique is being used to build the final model of news classification. This research shows that SVM with TF-IDF as feature performs well in further analysis. The accuracy of the prediction model is more than 90% having 52% correlation with the return label of a stock. This paper also proposes a real-time system which fetches news of any company on a real-time basis and displays its top five news and also predicts the adjusted close price of the next seven days.
Keywords: Text Mining, Human Sentiments, KNN, Random Forest, Multinomial Naïve Bayes, linear SVM, News.

1. Introduction

Indian stock market is unpredictable in nature. Numerous attempts have been made in order to predict the stock market movement. According to Behavioural economics, every individual is affected by psychological, cognitive, emotional, cultural and social factors on the economic decision. Is there any way to predict the stock market? There is a lot of information which affects the price movements of a stock, therefore, making it difficult for individual investors to predict the direction of stock price movements. Also monitoring such information on a real-time basis becomes even difficult. Nowadays, it is seen that stock prices not only depends on the historical factors and other macro-econometric variables, it also depends on the moods and sentiments of the investors that are being reflected in the news articles. Thus incorporating such opinions and sentiments can bring an improvement compared to the baseline model which is dependent only on the historical prices. In recent times the amount of news articles is increasing tremendously. Therefore analysing such news, extracting relevant information from it and using that information in prediction becomes an interesting problem. Our main objective is to build and test such a system which emphasises on the sentiment and author tone is incorporated within a financial news article in order to predict the better trend of the stock market in a short-term basis.

2. Related work

The paper of (Nguyen, Shirai, & Velcin, 2015)[5] showed an evaluation of the effectiveness of the sentiment analysis in stock prediction. They compared different sentimental models over 18 stocks for a period of one year to analyse which method yielded maximum accuracy. The first method, Price only method predicts the price of a stock by using time series analysis of historical data. The Second method is Human Sentiment where 15.6% of the messages in Message board dataset is being classified into five labels: Strong buy, Buy, hold, Sell and Strong Sell. SVM is used as a prediction model. The Third Method is Sentiment Classification, the remaining 84.4% of the messages which is not classified explicitly, a model is being built to extract the sentiments from those messages and then classified them into five categories. The feature representation is the bag of words from the title and content of the message and feature weightage is determined by TF-IDF and then fed into SVM classification model. The Fourth method is LDA (Latent Dirichlet Allocation) is used in order to extract the hidden topics. The LDA model is trained on the training dataset and unseen test set is being used to infer the topics by using Gibbs Sampling. After that the probability of each topic of each message is being calculated. Then these probabilities and the lagged prices are fed into the SVM prediction model. The Fifth method used is JST Model which extracts topics and sentiments simultaneously. Through Gibbs sampling 50 topics and 3 sentiments are being chosen. Next, the joint probability of topic and sentiment is calculated. Then these probabilities are integrated with the prediction model. The final and sixth method is Aspect Sentiment Model, in this model topics are extracted from the training dataset, topics occurring less than 10 times are being removed. Now based on the topic list the sentiment values are being extracted and opinion words are identified using SentiWordNet and sentiment scores are assigned: positivity, objectivity and negativity. Aspect sentiment model combined with the prediction model yielded the best accuracy.

The paper of (Li, Xie, Chen, Wang, & Deng, 2014)[4] made use of Harvard psychological dictionary and

Loughran-McDonald financial sentiment dictionary to construct a sentiment space. The model prediction accuracy and performance at different market classification is being compared. Each document is being represented as a vector of sentiment values by summing up the sentiment vectors of each word in the document. The news is collected from FINET. Two manual dictionaries is being used one Harvard Sentiment Dictionary (HVD) and Loughran McDonald Dictionary(LMD) for sentiment classification. SVM is used as text classification model. Another two approaches being used is SenticNet and bag of words. Sentiment Polarity in order to classify sentiments into 3 categories: positive, negative and neutral. By comparison from validation set it is being found that HVD, LMD, SN performs better than bag of words. Also HVD, LMD and BoW perform better than polarity approaches. After comparing all the approaches it is found that LMD is the best approach for finding the sentiments scores.

The paper of (Shri Bharathi & Geetha, 2017)[7] used a method RSS(Really Simple Syndication) for prediction in the stock market of Arab Bank Company. This RSS feed helps to collect news feed from the stock market which is being used as a dataset. Thus the system which is being developed automatically identifies the news opinions with the help of RSS news feeds and then predicts the stock movement. All the RSS feeds are then stored in an input sentence module as a document. Sentence splitting module cleans the news feeds and splits into parsed sentences. The NLP module is used to identify and extract subjective information from the source materials. Parts-of-Speech-Tagger is being used to read the text and assigns parts of speech to each word such as noun, adjective, verb etc. In Dictionary based approach is used to find opinion words and their polarities. It uses antonyms, synonyms and hierarchies in WordNet to determine word sentiments. Sentence Polarity module calculates the polarity of each sentence. If the polarity is positive then sentence is considered as positive and vice-versa. The polarity score value is being classified as positive, neutral and negative. If the opinion is positive then the stock goes up otherwise down. The polarity value for a month of 499 sentence is calculated which in then classified into positive, negative and neutral. Finally based on this it helps the marketers to buy or sell the stocks.

The paper of (Lee, Surdeanu, Maccartney, & Jurafsky, 2014)[3] have used 8K financial reports for all S&P 500 companies between 2002 and 2012 for their analysis of stock prediction. For Linguistic feature types unigram feature is being used and is lemmatized. Features are removed occurring less than 10 times and PMI is used to retain the linguistic features. Dimensionality reduction, NMF is implemented and then the resulting vector combined with baseline feature is fed into random forest classifier. The model is being trained with random forest using 2000 trees and compared with baseline model. The baseline model-1 is a deterministic system that predicts the system is UP when the actual earnings are better than expected. Baseline2 model uses 21 financial features, Unigram model uses unigram features in addition with 21 financial features. Then an ensemble model is being built based on the dimension of NMF. Using linguistics features improves the performance over non- linguistics features. Incorporating linguistics feature in the model improves the predictive power in short term. SentiWordNet lexicon is used to give scores to the words, positive words receives high score while negative words receives low score. Further Bigram and word clustering is used in order to combine two or more words. But bigram model does not improve the performance significantly over unigram model. Thus this unigram model improves short term prediction accuracy.

The paper of (Zhai, Cohen, & Atreya, 2011)[8] developed JAVA code and used the Stanford Classifier to quickly analyse financial news articles and using this model predict the S&P 500 index. Every article published in The New York Times from Jan 1987 to Jun 2007 used with proper annotation with date, category, and set of tags describing the content of the article. In order to classify natural language sentiment of news articles, two methods were tested for determining sentiment: manual and automatic ones using stock market results. Manual classification involved reading each article and assigning it a sentiment tag: positive, neutral, or negative. A class, NYT Manual Classifier, was built to aid in this process. Manual classification is time consuming. It was able to classify only for two months' worth of articles, January and June 2006 were chosen. January contains many articles summarizing the results of the previous year and speculating on the upcoming year. In June, journalists may be more focused on day-today movements of the stock market. Automatic approach used market movement in which log return: the log of today's close divided by yesterday's close is being used. System provides interesting analysis of market sentiment in hindsight, it is less effective when used for predictive purposes. The sentiment results produced could instead be an input to another trading system or simply be given to human traders to aid their judgments.

The paper of (Zhang & Skiena, 2010)[9] used Quantitative media (blogs and news as a comparison) data generated by a large-scale natural language processing and performed Comparative Study of Blogs and News, Large Scale Analysis and Sentiment Oriented equity Trading. They used Stocks listed in the New York Stock Exchange all 3248 stocks from period from 2005 to 2009. Media Data- Dailies, Twitter, Spinn3r RSS Feeds and LiveJournal (using Lydia Textmap.com). Their main focus was to find out Strength of correlation & Breakdown by Market Capitalization and Media Polarity vs. Stock Returns. They found out raw or derived blog/news variables are significantly correlated with some indicators in stock markets. Based on blog/news

sentiment data, it was designed a market-neural strategy, which is able to generate consistent returns for investors.

The paper of (Joshi, H. N, & Rao, 2016)[2] described about non quantifiable data such as financial news articles about a company and predicting its future stock trend with news sentiment classification. They used Apple Inc. Company's data for past three years, from 1 Feb 2013 to 2 April 2016. This data includes major key events news articles of the company and also daily stock prices of AAPL for the same time period. They created three different classification models which depict polarity of news articles being positive or negative. Observations show that RF and SVM perform well in all types of testing. Naïve Bayes gives good result but not compared to the other two. Experiments are conducted to evaluate various aspects of the proposed model and encouraging results are obtained in all of the experiments. If the news is positive, then it can be stated that news impact is good in the market, so more chances of stock price go high. And if the news is negative, then it may impact the stock price to go down in trend.

The paper of (Schumaker, Zhang, Huang, & Chen, 2012)[6] used Arizona Financial Text (AZFinText) system, a financial news article prediction system and pair it with a sentiment analysis tool. They used financial news articles from Yahoo Finance and represents them by their proper nouns as well as by the sentiment of the article. They made one software opinion finder in which if new document comes automatically three things will be come out Objective, subjective and Neutral(+/-/Neutral). The five verbs with the highest positive impact on stock prices are planted, announcing, front, smaller and crude. They believe that these results are attributable to investors reacting more strongly to negative articles.

3. Flowchart

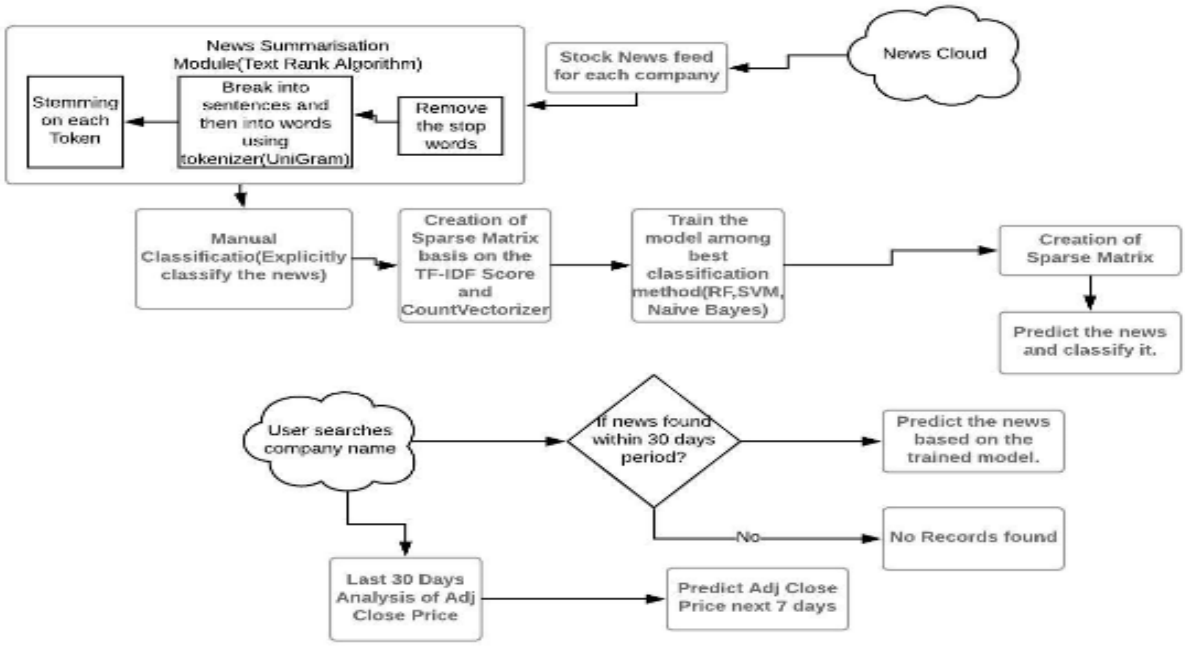

Figure-1: System design

4. Dataset

We have used two dataset for our analysis. The first is the historical price dataset and the second one is the news dataset

4.1. Historical Price Dataset

Historical prices are extracted from Yahoo Finance for the 37 stocks listed in Nifty50 and Nifty Next50 index from the year 2013 to October' 2018. For each transaction date Adjusted Close Price is extracted. (Nguyen et al., 2015)[5]The adjusted close price is the close prices adjusted according to the dividends and splits. Now from adjusted close price log returns are calculated by taking the log of today's close divided by yesterday's close. Now, we define the returns are labelled as positive for positive returns and negative for negative returns.

4.2. News Dataset

News for the same 37 companies enlisted in Nifty index is collected from Economic times for nearly six years from the year 2013 to October' 2018. For each date, company-specific news heading, description and news url

are extracted and stored in MySQL database.

5. Methodology

5.1. Semi Supervised Learning Technique

In semi supervised learning technique, both labelled data and unlabelled data will be used in order to create the model.

Figure-2 *Semi Supervised Learning Technique*

5.1.1. Manual Classification Method

This model is integrated with sentiments annotated by humans. (Zhai et al., 2011)[8]Manual Classification involved reading each article and assigning with a sentiment tag: positive or negative. Manual classification is time-consuming and we are able to classify around 10% of the news dataset. These sentiment label depicts whether the news is favourable to go ahead for the investment plan. Instead of using all the news, only the news annotated by the users explicitly has been used to train the prediction model. (Nguyen et al., 2015)Because the sentiments are annotated by humans, therefore, this features is used one of the strongest features to train the model. Some temporal diversity was desired because news situations can affect the market condition causing the overall sentiments to change. The news are manually classified in such a way so that it covers news for each year and for each company as well.

Some general properties have been used while classifying the news articles manually. A merger, joint ventures, technological investments, dividends are considered positive because it denoted companies have more cash on hand. While corruption, property seizing, lawsuits, sell of shares are considered as negative because it denoted more of negative aspects not favourable in terms of investment.

The news articles extracted after manual classification becomes one of the features of the machine learning model. For easier analysis, we have condensed each news articles using Text rank summarisation. Text rank is similar to page rank where each sentence is considered equivalent to a web page. The probability of going from one sentence to another is the similarity between the two sentences.

5.1.2. Prediction and Evaluation Component

(Deng, Mitsubuchi, Shioda, Shimada, & Sakurai, 2011)[1]Prediction and Evaluation component are for prediction and selection of the best model. Each summarised news articles is converted into a vector using Tf-idf vectorizer. From 10% of the news dataset that is manually labelled is divided into train and test set randomly in the ratio 8:2. Then the train set is subjected to four machine learning models Multinomial Naïve Bayes, KNN, Random Forest and SVM out of which linear SVM yield better performance in terms of accuracy which is then tested on the out of sample test set.

In the LSTM model, only one layer of LSTM and 10 hidden units is being used after hyper parameter tuning. The model is then compiled using 200 epochs, yielding the minimum validation RMSE.

5.1.3. Automatic Classification Method

Movements of the stock market were also used to generate classification. (Zhai et al., 2011)[8]We have used log return: log of today's close price divided by yesterday's close price. Positive returns are classified as positive while negative returns are classified as negative. It is seen there is a lag between the publishing of news and its effect on the stock prices. Typically most of the news has discussed the recommendation of buy and sale of stocks. Therefore analysing all the scenarios the labels of log return is used to compare the return labels with the manual classification of news articles. In addition, the market can be fickle and not always individual stocks may not respond to the news.

5.2. Price Only Method

In this method, only historical adjusted close prices are used to predict the stock movement. The purpose of this method is to investigate patterns in stock prices. (Nguyen et al., 2015)[5]In addition to this model is used as a baseline to evaluate whether there exists a correlation between the predicted return labels with the historical returns for every 37 companies.

The input features used in LSTM model is the historical adjusted closing price and output is the next day's price of lag length 1. The same model is executed for every 37 companies and tested on the out of sample data to study the variation patterns with the train data. After achieving a minimal loss through hyper parameter tuning we retrained the model on the whole data for each company and calculated the returns. On getting the return

values we labelled it as positive or negative depending on positive or negative returns. Now the predicted labels are compared with the original labels.

6. Results and Analysis

6.1. Experimental Setup

Around 21000 news articles is being extracted from economic times with respect to the chosen 37 companies, out of which 4500 is manually classified by explicit reading each news articles.

The manually classified news articles are randomly splitted into train and test set in the ratio 8:2. From our experimental analysis it is seen that both TF-IDF and Count vectorizer yielded the same result. We have preceded our further analysis with TF-IDF. Count vectorizer just counts the word frequencies but in TF-IDF the value increases proportionally to count but is offset by the frequency of the word in the corpus. Therefore TF-IDF vectorizer is used to convert each news articles into sparse matrix. Also it is seen from the literature review that TF-IDF is most commonly used method to convert a text into sparse matrix. Four machine learning models are fitted into the train set and tested on the out of sample data. SVM with TF-IDF gives better accuracy (Table - 1)

6.1.1. TFIDF Vectorizer

	Accuracy	
	Train	Test
Multinomial Naïve Bayes	0.74	0.708
KNN	1	0.75
RandomForest	0.677	0.669
SVM(Linear)	0.96	0.895

Table-1: *Comparison of Machine Learning Models using TF-IDF vectorizer*

Therefore it is seen that linear kernel of SVM model performs the best out of four machine learning models and gave .89 as f1-score.

6.1.2. Hypothesis

- $\mathbf{H_0}$: There is no relationship between sentiment of the news on a day and the return direction of the next day.
- $\mathbf{H_1}$: There is relationship between sentiment of the news on a day and the return direction of the next day.
- ***Chi-sq statistics***: 5.540
- ***p-value for Manual Label***: 0.018

		ReturnLabel	
		negative	positive
ManualLabel	negative	753	1438
	positive	670	1493

Table-2: *Confusion Matrix from Chi-square test.*

From the Chi-Square statistics we reject H_0 and conclude that there exists a relationship between the news on a particular day and the movement of return on the next day. Out of the several factors that affect the movement of stock returns, news sentiment has an impact of nearly 52.1%.

6.1.3. Chi-Square Test with respect to each company

- $\mathbf{H_0}$: There is no relationship between sentiment of the news on a day and the return direction of the next day company wise.
- $\mathbf{H_1}$: There is relationship between sentiment of the news on a day and the return direction of the next day company wise.

Company	p-value
Tata-Motors	0.04
PNB	0.09
HDFC Bank	0.2
Tata Steel	0.24
Wipro	0.12
Infosys	0.21
TCS	0.17
Asian Paints	0.02
Kotak Mahindra Bank	0.15
Indusland	0.19
Hindalco	0.03
Bajaj Finance	0.14
Sun Pharma	0.09
UltraTech Cements	0.09

Table-9: *Company wise Chi-sq significance Test*

From the above table we can see that news has more effect on the direction of movement of stock within 80% confidence interval. It is seen that mostly banking and IT sector is news driven and news has a strong effect on its stock prices.

6.1.4. Name Entity Recognition

We also found out frequently occurred verb and noun as part of speech tagging for negative and positive news.

6.1.4.1. Positive News

As per our analysis, some important verbs as 'launches', 'buy', 'raise' and 'rallies' and noun as 'Bank', 'Reliance' and 'Infosys' came out most occurred, so for further analysis this will impact more in order to classify positive news. So, Banking sector, Reliance and Infosys are the most profitable companies in last five years tenure.

6.1.4.2. Negative News

In Negative news frequently occurred verbs as 'stop', 'sell', 'seeks', 'slips' and 'asks' emerged, which shows negative trend of the stock.

6.1.5. Comparison of Return and Manual Label Positive and Negative Count by Year

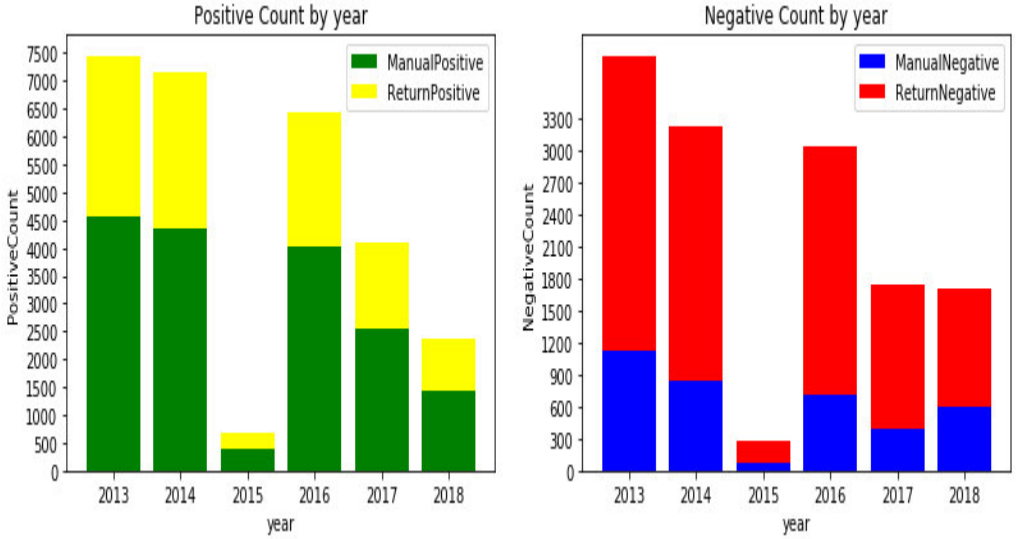

Figure-3: *Comparison of Return and Manual Label positive and negative Count by Year.*

From the first figure (left) it is seen that in the year 2013, 2014 and 2016, the count of manually labelled news as positive and return's movement in positive direction was nearly same whereas the count was least in the year 2015.Therefore the correlation between the news and the return direction was more compared to other years.

From the second figure (right) it is seen that in the year 2013, 2014 and 2016, the count of manually labelled news as negative is less compared to return's movement in negative direction whereas the count was least in the year 2015.

It is seen that the negative news have a persistent effect on the movement of return in negative direction. So we say that small number of negative news results in more downfall of return rather than more count of positive news driving the return in positive direction.

6.1.6. Comparison of Return and Manual Label Positive Count Company wise

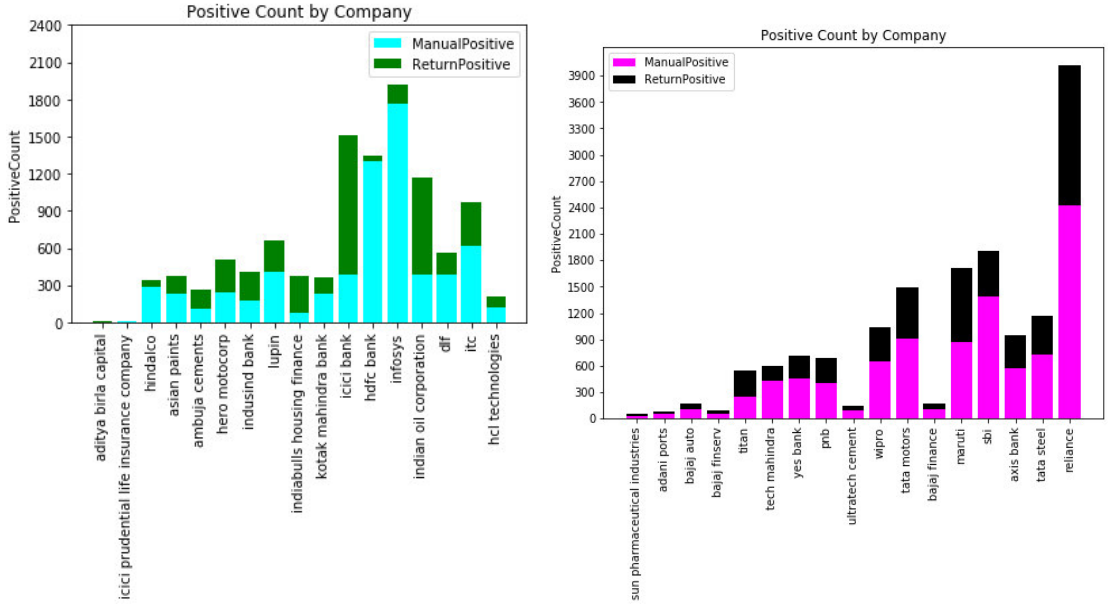

Figure-4: *Comparison of Return and Manual Label positive Count of 37 companies.*

From the above two figures we can conclude that the count of ICICI bank, HDFC bank, Infosys, Reliance, SBI and Maruti are among top positive counts. The count of return as positive and news as positive is unequal because more positive news failed to drive the returns in positive direction, this is due to the impact of positive news got diluted as a result of drop in index points of BSE Sensex and Nifty and also due to industry impact as well.

6.1.7. Comparison of Return and Manual Label Negative Count Company wise

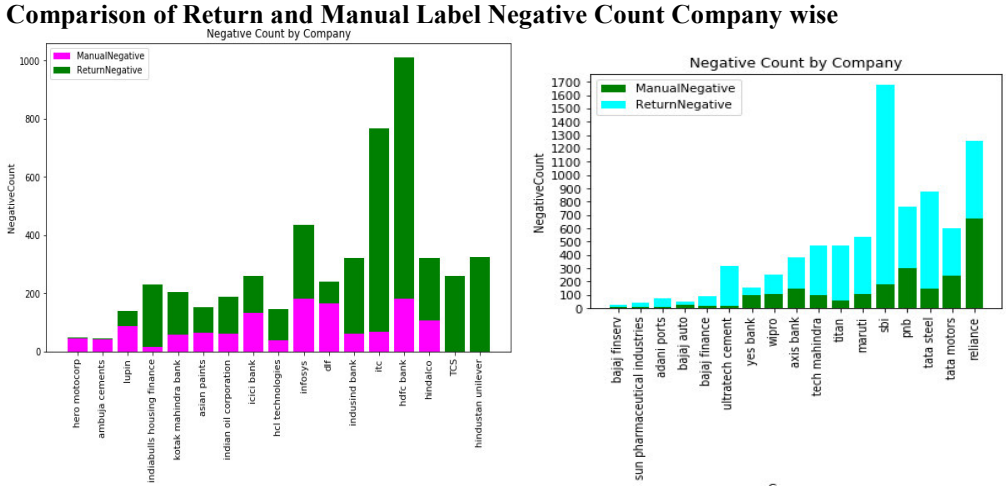

Figure-5: *Comparison of Return and Manual Label positive Count of 37 companies.*

From the above two figures we can conclude that the negative count of SBI, PNB, Reliance, ITC, and HDFC bank are among top negative counts. It is also seen that the count of return as negative and news as negative is unequal, this is due to the fact even though less news are negative, the direction of movement of return is more in negative zone. From this one important inference can be drawn that negative news have very long impact on stock market. From the analysis it is found that the impact of news got diluted because of volatility of BSE Sensex and Nifty index. This irregularity of news and direction of return movement was due to corruption, huge fall in banking sector because of PNB scandal, rise in petrol taxes and many more.

Example which shows the impact of negative news to the stocks, on 16, April, 2016 news came against TCS, “US court fines TCS $940m for info theft”. Due to this news for next four days TCS stock felt down drastically.

6.2. Real Time News Prediction

In this approach we built an interface in which user enters the symbiotic notation of any Indian companies, the system will display top five news from that point of time to last 30 days classified as positive or negative. Also the system will provide prediction of Adjusted Clos Price that stock for next seven days from the current date by

fetching data from Yahoo Finance using LSTM. Therefore by analysing all these prospects it is left with the user to decide on his/her investment plan.

7. Conclusion and Limitations

Stock price prediction is a challenging task because the stock returns are being affected by many factors. This paper presented a real time prediction model to integrate the sentiment of news articles for the prediction of stock return movement either in positive or in negative direction. This research proposed two models one is to extract the sentiments based on the sentiments of news articles, second is the prediction of movement of stock returns based on adjusted close price. While news articles remain a useful source of information for the determination of market sentiment but they are difficult to analyse. Furthermore the success of our model largely depends on the exploitation of market inefficiencies.

Even though our system provides an interesting analysis of news sentiments, it is less effective for prediction purpose. This is due to the fact that out of several factors affecting stock price movement impact of news only constitutes 52% correlation. Our prediction model has failed to capture the market driven news and the impact of macro-economic variables. Also our model has not considered transaction cost, borrowing cost and the tendency of large trading to move the market. Therefore our model is insufficient for prediction of large movements in stocks.

8. Future Scope

There are a number of ways this work can be improved. One idea is to include economic indicators as one of the feature to integrate with our model and evaluate the performance. The effect of macro-economic variables can be included to see its impact. Moreover we can identify some patterns in the news articles, cluster them and see its correlation with return so that more weights can be given to such news. This will allow us to model the relationship between news sentiment and market performance in a better way. For traders and user perspective a web based application can be developed where user can select a list of companies and see their movement of stock returns with respect to news which will help them in deciding their invest plan.

9. References

1. Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., & Sakurai, A. (2011). Combining technical analysis with sentiment analysis for stock price prediction. *Proceedings - IEEE 9th International Conference on Dependable, Autonomic and Secure Computing, DASC 2011*, 800–807. https://doi.org/10.1109/DASC.2011.138
2. Joshi, K., H. N, B., & Rao, J. (2016). Stock Trend Prediction Using News Sentiment Analysis. *International Journal of Computer Science and Information Technology*, *8*(3), 67–76. https://doi.org/10.5121/ijcsit.2016.8306
3. Lee, H., Surdeanu, M., Maccartney, B., & Jurafsky, D. (2014). On the Importance of Text Analysis for Stock Price Prediction. *Lrec*, (2009), 1170–1175. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/1065_Paper.pdf
4. Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*(1), 14–23. https://doi.org/10.1016/j.knosys.2014.04.022
5. Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052
6. Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458–464. https://doi.org/10.1016/j.dss.2012.03.001
7. Shri Bharathi, S., & Geetha, A. (2017). Sentiment Analysis for Online Stock Market News Using Rss Feeds. *International Journal of Current Engineering and Scientific Research*, *4*(4), 58–63. Retrieved from http://troindia.in/journal/ijcesr/vol4iss4/58-63.pdf
8. Zhai, J. J., Cohen, N., & Atreya, A. (2011). CS224N Final Project : Sentiment analysis of news articles for financial signal prediction. *Mimeo*, 1–8. https://doi.org/10.1016/j.dss.2008.04.001
9. Zhang, W., & Skiena, S. (2010). Trading Strategies to Exploit Blog and News Sentiment. *Proceedings of the Fourth International AAAI Conferenece on Weblogs and Social Media*, *34*, 375–378. https://doi.org/10.1016/j.jbankfin.2009.11.025