IOSR Journal of Economics and Finance (IOSR-JEF)
e-ISSN: 2321-5933, p-ISSN: 2321-5925.Volume 10, Issue 3 Ser. I (May. – June 2019), PP 74-81
www.iosrjournals.org

Impact of Company News and Consumer Sentimentson Stock Price Fluctuation

DrSayantani Roy Choudhury[1],Anagha Karanam[2]
[1](Faculty, Economics & Statistics Area, Praxis Business School, Kolkata, India)
[2](Post Graduation Student, Praxis Business School,Kolkata, India)

Abstract:*Stocks fluctuate because of many economic and non-economic reasons and many researches tried to quantify reasons for fluctuations in order to predict how much a stock will rise/fall. One of the most important factors is consumer sentiment and news. In this research, intensity-based sentiment analysis is used to measure sentiments of news articles and social media(tweets)of considered stocks. Regression Analysis on the sentiment scores is obtained which reveals that there might be a strong correlation between news, social media & stock price fluctuation for some stocks and thus quantifying their sentiments is essential to study stock fluctuations with precision.*

Keywords: *ANOVA, Regression Analysis, Stock fluctuation, Sentiment Analysis, VADER*

Date of Submission: 03-05-2019 Date of acceptance: 17-05-2019

I. Introduction

Stock price of any company depends on various factors like the following-

1. Company news and performance
2. Industry performance
3. Investor sentiment
4. Economic factors

But many of the sub-factors within the above four are interrelated. Industry performance can definitely get influenced by the Economic factors. Again, company performance depends on industry performance. Investor sentiment is the most complicated factor among all. It is itself a function of all the three other factors. Investor sentiment or confidence can cause the market to go up or down, which can cause stock prices to rise or fall. The general direction that the stock market takes can affect the value of a stock. A strong stock market where stock prices are rising, investors get the confidence to invest. Stock price rise is tied to economic recovery or an economic boom, as well as investor optimism. A weak market where stock price is falling, can reduce investors' confidence. It can reduce investment and as a result, stock price. Therefore, consumers' sentiment is a cause of as well as an effect on the fluctuations in stock prices. It seems to be the most important factor behind stock price fluctuations.Therefore, we focused on consumers' sentiment in our paper. We tried to find out the association between company news with consumers' sentiment, and then, association between consumers' sentiment with stock price fluctuations.

II. Literature Review

We can divide the literature review in two portions:

a. Literature related to factors behind stock price fluctuations
b. Literature related to text analysis

a. Literature related to factors behind stock price fluctuations

There has been huge amount of literature present where it has been identified that macro as well as micro variables are significant in explaining the fluctuations in stock prices. Homa and Jaffee (1971) showed that there is significant association between stock prices and money supply. Rudolph (1972) established a statistically significant relationship between money supply and stock prices (SP) where SP included 500 poor and standard stock indexes. Nelson (1976) used data of the period 1953- 1994 and established an inverse relationship between stock returns and expected as well as unexpected inflation for US economy. Abdalla and Murinde (1997) checked in financial markets of different countries like India, Pakistan, Korea and Philippines. They found out that there exists a one-way association between exchange rate and stock prices. Park and Ratti (2008) did a regression analysis taking US and other 13 European economies and after

considering the data for 1986-2005 showed stock prices are both negatively and positively related with oil prices depending upon the fact whether a country is oil importing country or oil exporting country.
Mukherjee and Naka (1995) used VECM (vector error correction model) on Japanese stock returns and macroeconomic variables and found that stock market was co-integrated with the six dependent variables. Mock (1993) verified that there was a unidirectional causality between interest rates and closing stock prices using ARIMA approach and Granger Causality tests. Weihong Huang and Yu Zhang (2014) examined stock price fluctuation asymmetry, the asymmetry between stock price rise and fall speed. Based on inverse statistics, a new measurement, named as asymmetry index, was proposed to evaluate this asymmetry. They calculated and compared asymmetry indices of historical prices from ten stock markets. It was found that in most stock markets, price fall was faster than price rise; while in China and India, price rise was generally faster than price fall. Arpit Bhargava, Ankush Bhargava, Surbhi Jain (2016) studied the relationship between macro variables such as Inflation, Index of Industrial production(IIP), Money Supply, Oil prices, Exchange rates, Gold prices and Gross domestic product (GDP) and Stock Prices using time series regression. The sample was for the period 2004-2013 on quarterly basis. The study revealed that only Exchange Rate, Oil Prices and Inflation have significant impact over Stock prices.

b. Literature related to text analysis

Several attempts to classify the sentiments present in the text have been done in various sectors such as movie reviews, stocks, recommender systems, product reviews etc. One of such important attempts is to predict the movement of stocks using the sentiments from news and social media. AnushGoel and Arpit Goel (2011) discuss the movement of stock price using Daily Score Computation and Score Mapping on twitter data. In the paper, 'which news moves stock prices? A textual analysisby Jacob, Ronen, Shimon and Matthew (2013), an attempt has been made to identify the relationship between stock prices and news. News of each day is first labelled into a group based on bag-of-words concept and a daily score is computed using a standard method of summing the positive and negative words. $S = (P - N) \div (P + N + 1)$, where P and N stand for the number of positive and negative words, respectively. The algorithm then checks if stock moves and if there is relevant news labelled on a given day and also analyses the movement as positive or negative using tone of text.

Although the concept is very clearly stated, the algorithm/methodology used is obsolete in today's world of social media and micro-blogging. These algorithms will not work to identify social media text where lot of information is aggregated in a sentence or two. Also, older lexicons fail to include emoticons and abbreviations which make up a major part of understanding the sentiment of the text.

In 'Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty', Aditya Bhardwaj, Yogendra Narayan, Vanraj, Pawan, Maitreyee Dutta (2015), different machine learning and lexicon based approaches were presented to analyse the sentiment in a given text for Indian stock scenario.

As stated by Alexandra Balahur and Ralf Steinberger(2009), The European Commission's (EC) Joint Research Centre (JRC) has developed a number of news analysis systems. Called the European Media Monitor (EMM), EMM scrapes more than 2200 news sources every day, providing latest news in standard RSS format. EMM also classifies similar news together. Although it is widely used, it fails to consider and compare public opinions on the articles it delivers. An attempt like EMM along with social media sentiments has not been successfully implemented in India yet. This paper explores such an attempt.

VADER (Valence Aware Dictionary for Sentiment Reasoning) by C. J. Hutto, Eric Gilbert (2014), is an approach to classify given text into sentiments considering valence (intensity) of the text. The underlying generalized, valence-based, human curated gold standard sentiment lexicon was built to be made suitable for today's social media, micro-blogging text which convey a lot of information in a single line using abbreviations, emoticons, punctuation marks etc. Built on top of other famous sentiment analysis algorithms such as SentiWordNet, (Baccianella S., Esuli A., and Sebastiani F. (2010)). In Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis by Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018), the authors focussed on volume of tweets along with the sentiments with an assumption that people generally tweet optimistic about Bitcoin. Nevertheless, VADER was used to analyse sentiments on tweets collected about Bitcoin to decide whether to buy or sell.

III. Research Question

A large number of research attempts have been done to explore the reasons behind stock price fluctuations. But, still, it is an important topic to be researched. There are huge numbers of factors behind the ups and downs of the stock prices.

Our research question is to find out whether there is an association between the stock price fluctuations of a company and news related to that particular company appearing in any newspaper and social media.

The model
Stock price fluctuation of a company= a + b (scores generated from the published news on that company)

Hypothesis
H0: b= 0
H1: b> 0

IV. Methodology

We divided the methodology portion into two parts.

a. In the first module, after collecting data from twitter, news and google search results, we calculated compound valence based sentiment analysis scores for each data source.
b. We used regression analysis to understand whether there is any connection between stock price fluctuations and the scores which we obtained in the 1st part.

To test our hypothesis, we considered three different stocks across different sectors as shown in Table 1.

Table 1: List of stocks considered for testing the hypothesis

HDFC Bank	Reliance Industries	Tata Steel

4.1 Dataset

Data is collected from three different sources as follows:

4.1.1 Twitter Data

Publicly available twitter data from January 1st 2018 to December 31st 2018 is scraped. The scraped data includes username, tweet-id, timestamp, url of the tweet, likes, replies, retweets, text of the tweet and html in the tweet if any. The problem with using the famous twitter API is that it can scrape tweets only for the last 7 days. So, taspinar'stwitterscraper package is used. The python package is installed. The query in twitter scraper to fetch tweets returns tweets in multiples of 20. So, if the start and end date are given in the range of 1 year, and if the maximum number of tweets to fetch is given as 5000, the output does not have uniform number of tweets for all days/months.

In order to solve this problem, the same query is written by considering the time frame in groups of 4 months each and three csv files are separately generated. Table 2 shows the number of rows obtained for each of the three quarters for different stocks. As observed, there is a lot of variation in the number of tweets obtained for considered stocks.

Table 2: Number of rows obtained for each of the three quarters for stocks considered

Time Range	HDFC Bank	Reliance Industries	Tata Steel
January - April	3109	3562	240
May - August	3356	4357	343
September - December	3682	4524	374

4.1.2News and Newspaper Data

News and Newspapers have important micro economic information about the company, they state facts about the company which can cause the stock price to fluctuate. In this section three famous news sources in India are considered which are Economic Times, The Hindu. and NDTV. News archive is scraped for the given date range (January 1st 2018 to December 31st 2018) for urls of news articles which contain the company name.

For example: The Hindu maintains an archive containing links to daily news and those urls which have the word 'hdfc bank' are fetched.

In order to fetch relevant urls, a set of urls are first studied and a regular expression is written which fits all the possibilities. For example: Urls of reliance industries can also contain *ril* apart from the word *reliance-industries*.

For each of the three sources the following steps are performed:

1. The archives are scraped, to find urls of the articles which match the regular expression.
2. Using the "beautiful soup(bs4)" package, content from these urls is then fetched in a csv file.
3. The final csv file for each news source contains the timestamp, text of the article and url.

Table 3 shows number of articles for tata steel which were fetched from different news sources.

Table 3: Number of articles fetched from different news sources for Tata Steel

Source	Number of articles fetched
Economic Times	285
NDTV	27
The Hindu	34

4.1.3 Google Search Data

Micro events in a company can also influence the price of a stock. These events might not be covered in the daily newspapers but can definitely be found using a web search. To illustrate further, either because of its less importance or space limitations, newspaper and news may let go some news. But, this information could be important for predicting movement of a stock price. To consider this window of possibility, google search results for news dated between January 1st 2018 to December 31st 2018 is obtained and sentiment analysis is performed.

To make obtaining data easier, top five websites which continuously publish news articles are considered which are livemint.com, business today.in, news18.com, reuters.in and moneycontrol.com.
For each of the given websites, the following steps are performed to obtain the data:

1. A python code is written to perform a google search and fetch articles in the given date range. For example: the search string for businesstoday.in is created as: site:businesstoday "hdfc bank"
2. A piece of beautiful soup code written scrapes the google search results and fetches the urls.
3. The urls are then scraped to fetch the text from the articles.
4. Five csv files for each of the websites is obtained which contain small news which might affect the stock price.

4.2 Data Pre-processing

Data collected from different sources has to be pre-processed before feeding it to the sentiment analysis algorithm.

4.2.1 Twitter Data

As shown in Fig 1, a python script is written to read the obtained three csv files in three different data frames. Columns which are not required are dropped and only the timestamp and text of the tweet is retained. An outer join on these three data frames then combines the data into one single data frame.
Finally, a single data frame containing all the tweets is obtained. It is important to note that there can be multiple tweets with the same date. It is retained this way at this stage and later averaged when date is made the index of the data frame. The timestamp variable is converted to Date type in python and strftime() function is used to strip the time values and change the date format to dd-mm-yyyy.A single csv file for a stock with date and text is obtained on which sentiment analysis can be performed.

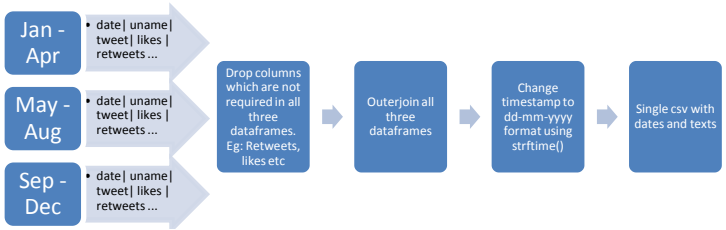

Fig 1: Combining three separate data frames into a single data frame.

4.2.2 Newspaper, News Data

1. Each of the three chosen news sources has the text,timestamp and url in a csv file. They are read into separate data frames, the url column is dropped and an outerjoin is performed.
2. The timestamp variable is converted to Date type in python and strftime() is used to strip the time values and change the date format to dd-mm-yyyy.

Similar to what was said about twitter data, there can be multiple rows for the same date. A single csv file for a stock with date and text is obtained.

4.2.2 Google Search Data

The process used for pre-processing Google search results is similar to News Data. The same method is applied on five csv files obtained from five news sources.

As an end result, there are 3 csv files (twitter, news and google search) for each stock with date and text. In the next step Valence based Sentiment Analysis is performed on the text present in these csv files.

4.3 Sentiment Analysis

4.3.1 VADER (Valence Aware Dictionary for Sentiment Reasoning)

One of the most curious research areas in Natural language processing (NLP) is Sentiment Analysis. Extensive research is going on this area and many solutions are floating around, the accurate way to classify text, especially social media data is still a mystery.

A lot of sentiment analysis approaches (algorithms) depend on underlying sentiment lexicon(dictionary). A sentiment lexicon is just a list of words which are already labelled. But, using these pre-existing lexicons to perform sentiment analysis does not suffice because there is a level of intensity associated with the word. For example, ranking "happy" and "ecstasy" with an intensity scale would be more accurate than ranking them both as positive.

Explained by C. J. Hutto, Eric Gilbert (2014),sentiment lexicons are broadly classified as two types:

1. Semantic Orientation(Polarity-based) Lexicons
2. Semantic Intensity(Valence-based) Lexicons

There are about eleven widely used techniques in these two categories. Examples include LIWC, GI, Hu & Liu in Polarity based and ANEW, SentiWordNet in Valence Based. However, all these have their own shortcomings, the important shortcoming being challenges applying to social media and microblog data.

Use of symbols (emoticons), abbreviations and expressing a lot of information in a few linesmakes understanding the sentiment accurately a tough task. C. J. Hutto, Eric Gilbert (2014), calls this problem contextual sparseness. Using Machine Learning Approaches for sentiment analysis is a daunting task because there is usually not enough text data available to train and test. Moreover, performing training on text data is computationally expensive.This is where VADER serves the purpose to an extent. VADER (Valence Aware Dictionary for sentiment Reasoning) was created by C. J. Hutto, Eric Gilbert (2014).

VADER considers five rules that embody grammar and syntax conventions in order to effectively express the sentiment of given text. They are punctuation, capitalization, degree modifier, contrastive conjunctions and trigram preceding the text.It calculates and groups text into 4 different scores, positive, negative, neutral and compound. So, we not only get if the score is positive or negative, we get the intensity of the score. The score ranges between -1 to +1. So, VADER gives you not just that the text is positive, but also how positive it is.

For example, "The best I can say about the movie was that it was interesting". Here, giving a positive sentiment just because there is the word, interesting does not make sense. VADER gives a compound score based on the intensity of the words surrounding 'interesting' and also considers the change in tone of the sentence.

VADER's correlation coefficient shows that it performs as good as human raters. Moreover, as verified, the classification accuracy of VADER is actually better than individual human raters. VADER performs well with emojis, acronyms, microblogs. It can consider multiple sentiments at once.(figurative speech) And, the best part is that the sentiment lexicon has been validated by humans.VADER has been tested against some Machine Learning approaches and it is known to perform equally well, or even better. Moreover, it is simple to implement than training from a huge set of training data

We use VADER to perform sentiment analysis on our pre-processed data sets.

4.3.2 Application

For each of the csv files from three different sources, the following steps are performed:

1. VADER offers a polarity scores () function which calculates the valence based sentiment scores. This function is called for each row of the dataframe.
2. The text column in each row is dropped and only the four scores(positive, negative, neutral and compound) is appended to each row containing the date.
3. A groupby with mean for all the scores for the rows with the same date gives a unique row for each date.
4. Making date the index and Re indexing for the missing dates gives a final dataframe which has four columns of scores for each date and NaNs for those dates where there no sentiment scores.
5. The final data frame is exported to a csv file.

The process is shown in Fig 2. The final result consists of three different csv files containing sentiment scores, calculated from three different data sources i.e. twitter, newspapers and google search.

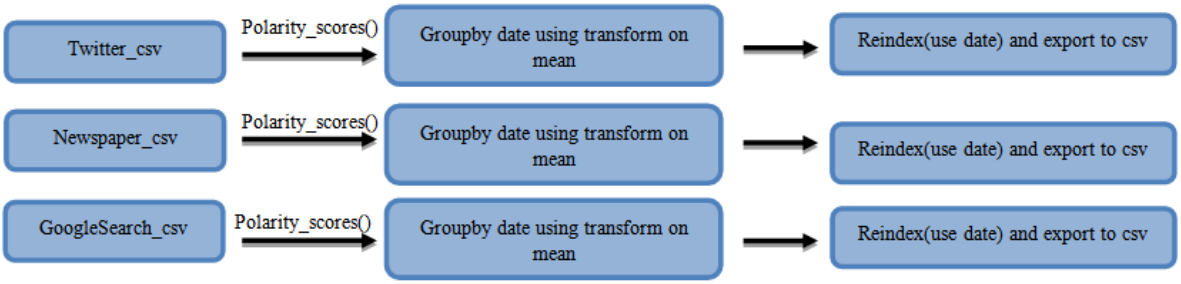

Fig 2: Applying VADER on the pre-processed data for the three data source

4.4 Regression Analysis

In the second module, based on the assumption that the effect of a news on the movement of stock price is usually applicable for next 5 days, regression analysis is applied to check if stock prices change based on last 5 day moving average of the obtained sentiment analysis scores.

Stock price fluctuation of a company= a + b (sentiment analysis scores) (1)

Regression has been done on this above model with three sets of daily data of the year 2018 for three different companies; Tata Steel, Reliance Industries and HDFC Bank. From 6.1 we got the sentiment analysis scores of the three companies. Stock prices for each of the three companies have been regressed on the last 5 day moving average of the obtained sentiment analysis scores.

V. Results

Following tables show the results of the regression analysis. ANOVA tables show the overall model fit and the other table shows the significance of association between text score on stock prices.

Table 4: TATA Steel

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	160807.835	1	160807.835	11.715	**.001***
	Residual	3363152.144	245	13727.152		
	Total	3523959.979	246			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	971.726	24.847		39.108	.000
	text_r	43.466	12.700	.214	3.423	**.001***

Table 5: Reliance Industries

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	8753.873	1	8753.873	2.491	**.116**
	Residual	857608.983	244	3514.791		
	Total	866362.856	245			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	617.564	9.719		63.539	.000
	text	-16.339	10.353	-.101	-1.578	**.116**

Table 6:HDFC Bank

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	37285.545	1	37285.545	4.004	**.047***
	Residual	2281641.059	245	9312.821		
	Total	2318926.605	246			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	1993.897	13.140		151.743	.000
	text	**19.389**	9.690	.127	2.001	**.047***

The results show there is a significant relationship between stock price and scores generated through the published news in case of Tata Steel and HDFC Bank (<0.05). Therefore, in these two cases, H1: b> 0 is true. That implies there is a statistically significant direct association between text scores and stock prices.
For reliance industries, the relationship is insignificant (.116). Thus, for this company, H0: b= 0 is true. Therefore, the ups and downs of stock prices of Reliance Industries are not significantly influenced by the sentiment/text scores.

VI. Conclusion

The volatility of stock prices is a complex function of many influencing factors. One factor is definitely the 'consumers' behavior'. There is some obvious impact of the different published news about a company on the stock purchasing behavior of a consumer. For HDFC Bank and for Tata Steel, this impact is significant on the purchasing behavior and as a result it influences the stock prices of these two companies. The association between these two is positive and this implies stock prices go up when test scores are high. On the other hand, in case of Reliance Industries this factor is weak; it is not significantly influencing the purchasing behavior of the consumer as well as the movement of stock prices.

Although we cannot make a strong statement yet about 'if and how much' would a stock fluctuate when it appears in news and social media, we should definitely consider sentiments of news and social media as one of the important influencing factors when performing analysis. By this one can predict the movement of stock prices with more precision.

References

[1]. Arpit Bhargava, Ankush Bhargava, Surbhi Jain (2016); Factors Affecting Stock Prices in India: A Time Series Analysis; *IOSR Journal of Economics and Finance (IOSR-JEF), e-ISSN: 2321-5933, p-ISSN: 2321-5925.Volume 7, Issue 4. Vol. I (Jul. - Aug. 2016), PP 68-71.*

[2]. Issam Abdalla and Victor Murinde (1997); Exchange rate and stock price interactions in emerging financial markets: evidence on India, Korea, Pakistan and the Philippines*; Applied Financial Economics, 1997, vol. 7, issue 1, 25-35.*

[3]. J. Allan Rudolph (1972); TheMoney Supply and Common Stock Prices;*Financial Analysts Journal**; Vol. 28, No. 2 (Mar. - Apr., 1972), pp. 19-25.*

[4]. Kenneth E Homa and Dwight M Jaffee(1971); The Supply of Money and Common Stock Prices; *Journal of Finance, 1971, vol. 26, issue 5, 1045-66.*

[5]. Nelson C (1976), "Inflation and Rate of Returns on Common Stocks*", Journal of Finance, Vol. 31, pp. 471-483.*

[6]. Park Jungwook and Ronald A. Ratti (2008); Oil price shocks and stock markets in the U.S. and 13 European countries; *Energy Economics; 30 (5): 2587-608.*

[7]. Tarun K. Mukherjee, Atsuyuki Naka (1995); Dynamic Relations between Macroeconomic Variables and the Japanese Stock Market: an application of a vector error correction model; *The Journal of Financial Research, volume18, issue2, Pages 223-237.*

[8]. WeihongHuangand Yu Zhang (2014); Asymmetry Index of Stock Price Fluctuations; *Journal of Global Economy; Vol. 2(3): 118*

[9]. Anshul Mittal and Arpit Goel(2011); *Stock Prediction Using Twitter Sentiment Analysis; Stanford CS229 projects.*

[10]. Jacob Boudoukh, Ronen Feldman, Shimon Kogan, Matthew Richardson(2013), Which news moves stock prices?A textual analysis; *National Bureau of Economic Research, Inc. ; NBER Working Papers 18725.*

[11]. Aditya Bhardwaj, Yogendra Narayan, Vanraj,Pawan, Maitreyee Dutta(2015);Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty; *Procedia Computer Science, Volume 70, pp. 85-91.*

[12]. Balahur, Alexandra & Steinberger, Ralf & van der Goot, Erik &Pouliquen, Bruno &Kabadjov, Mijail. (2009); Opinion Mining on Newspaper Quotations; *Web Intelligence Consortium, 523-526. 10.1109/WI-IAT.2009.340.*

[13]. Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018); "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis"; *SMU Data Science Review: Vol. 1: No. 3, Article 1.*

[14]. C. J. Hutto, Eric Gilbert(2014); VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text; *International AAAI Conference on Web and Social Media.*

DrSayantani Roy Choudhury and Anagha Karanam. "Impact of Company News and Consumer Sentimentson Stock Price Fluctuation." IOSR Journal of Economics and Finance (IOSR-JEF), vol. 10, no. 3, 2019, pp. 74-81.