International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 8– No.7, May 2015 – www.ijais.org

Empirical Study of Relationship between Twitter Mood and Stock Market from an Indian Context

Saurav Kumar
M.Tech, Final Year
Dept. of Computer Science Engineering
IEM, Kolkata
West Bengal, India

Siddartha Maskara
B.Tech, Final Year
Dept. of Computer Science Engineering
IEM, Kolkata
West Bengal, India

Nitin Chandak
B.Tech, Final Year
Dept. of Computer Science Engineering
IEM, Kolkata
West Bengal, India

Saptarsi Goswami
Asst. Professor
Dept. of Computer Science Engineering
IEM, Kolkata
West Bengal, India

ABSTRACT
Various studies have been conducted to investigate relationship between sentiment from investors or from news and stock market movement. From literature study it is observed, there is no systematic study conducted on the same for India, which is one of the leading emerging markets of the world. In this paper, 'twitter' have been used as the source of the news as mostly all popular channels publishes news through tweets. Corpora of 0.3 Million tweets have been collected between July 2014 to Mar, 2015, from 30+ relevant twitter handles. The polarity of the news has been extracted and shown to have a significant correlation with stock market movement measured in terms of 'Sensex' and 'Nifty', the major stock indices of India. Relationship of the sentiment with other macroeconomic factors like Gas and Oil Price, Exchange rate etc. has also been examined.

Keywords
Stock Market Prediction, Mood, Sentiment Analysis, Sensex, Correlation Tweets

1. INTRODUCTION
Every day on internet, as of 2014, 294 Billions emails are sent, 3.5 billion Google searches are done, 3.5 Billion Facebook messagesare posted, 500 Million Tweets are shared. This is a staggering amount of data created each day on the internet, which is unstructured and noisy yet represent a valid source of information. Twitter which is a popular micro-blogging site is one of the main sources of information.

The information from such internet, social media sources can be applied to many problems. It is reshaping healthcare [1] [2], the way business is done [3]. Social media has given birth to new subjects like 'Socialnomics' [4].One such interesting problem is predicting the stock market movement based on sentiment from Tweets. Over last few years researchers have used various methods to try and find correlation between the stock markets.

Earlier research on prediction of the stock market was based on random walk theory and Efficient Market Hypothesis [5] which was not very successful as there were anomalies [6]. According to EHM, news is the most important factor affecting the stock market rather than present and past prices. But news itself is so unpredictable that EMH cannot predict with more than 50% accuracy [7].

Although a lot of research work has gone into twitter sentiment analysis and using it to predict stock market movements in countries like USA [8], UK, France, China and other countries. However from literatures study it was observed, there is no systematic study done for Indian stock market to establish a relationship between the stock indices movement and sentiment extracted from news.

In this paper, it has been studied

- Relationship of Nifty and Sensex with the sentiment extracted from news media through Twitter on a dataset consisting of 0.3 Million tweets collected over 9 months.
- Also as general social media is noisy, selected twitter handles, related to finance news has been identified and used. Further details are available in Section 3.
- 31 twitter handles of various financial news media and investment farms were considered.

The organization of the rest of the paper is as follows. An insight about Twitter and its importance as data source is given in Section 2. Section 3 presents the steps of methodology executed in this research work. Followed by the results and analysis of the correlation achieved in Section 4. Section 5 contains suggestions for future research along with concluding remarks.

2. TWITTER AS DATA SOURCE
Twitter, the microblogging social networking website, ranks as the 9th most popular website according to Alexa rankings as of January 2015, having around 284 million monthly active users. Everyday around 500 million such tweets are made, most of which are publically available.

The popularity of twitter among academicians and research scholars is increasing as collecting data based on user's predefined parameters using "The Twitter Streaming API" is easy, unlike other social media websites like Facebook from where retrieving data is a tough task. The "Twitter Streaming API" is a capability provided by Twitter that allows anyone to retrieve at most a 1% sample of all the data. Morstatter suggested that for large datasets, or big issues that generate lots of traffic, the 1% is apparently 'faithful' to the full stream, with a common set of top keywords and hashtags [9]. As the collected tweets are over 9 months of duration and has significant number of tweets, the general trends observed over the sample is expected to hold over entire dataset.

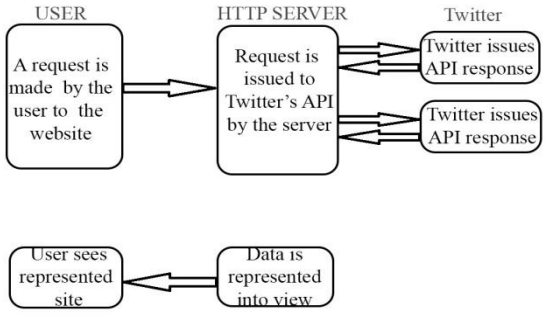

Fig 1: User-Client Request-Response

Real time messages from Twitter have allowed researchers to work on various fields like prediction political election result [10] [11], political sentiment [12] [13], predicting box office return [14], predicting stock market [15] [16] and many other. Hashtags which are used to annotate the tweets also have been used in many research works [17] [18] [19] [20] [21].

3. EXPERIMENTAL SETUP

The tweet extraction process on the Twitter usually extracts different numbers of tweets per day. Each tweet has many attributes, only few of them that are used by researches

- **Text** – The text of the tweet
- **ScreenName** - Screen name of the user who posted the tweet
- **ID** – unique ID of the tweet
- **Retweet** – 'True' if the tweet is retweeted by some other user else 'False'
- **Retweet_count** – The number of times this status has been retweeted
- **Favorited** - Whether this status has been favorited
- **Truncated** - Whether this status was truncated
- **Geo** – Returns 'null' if tweet is not geo-tagged otherwise return the location from where it is tweeted.
- **Created** – Timestamp at which tweet was posted.

Many of the attributes return value 'null' like favorited, retweet, geo. So for each tweet only the following few attributes have been extracted as shown in table 1:-

Table 1. Collected Fields of the tweets

Identifiers	Example
Id	500612990433906000
Screen Name	@EconomicTimes
Timestamp	8/16/2014 14:30
Tweet	Why FDI may be the best option to revitalize domestic defence production http://t.co/Kw6RoWfEC9
IsRetweet	FALSE

3.1 Data Collection

The tweets have been collected as a weekly process. Over a period of 9 months, July, 2014 – March, 2015 only publicly available tweets were collected (around 309,124 tweets).

Identifying the Twitter accounts which can affect the general sentiment of the public was an important step. Extensive study has been carried out to select appropriate Twitters handlers. For which some basic steps were followed:-

1) Initial search was done based on keywords like '#Sensex', '#NIFTY', '#economic' etc.
2) It helped us to identify that which accounts to usually tweet in large volume using these keywords.
3) Out of which 31 accounts with rich diversity were selected as shown in table 2.
4) From the selected twitter handles , data was extracted on daily basis
5) As the 'Id' of the tweets are stored an appropriate incremental process for daily extraction could be set up easily.
    - The maximum twitter id for the handle is retrieved from the stored tweets
    - Only those tweets which have an 'Id' greater than the mentioned Id in previous step are extracted.

Table 2. Some of the selected Twitter accounts

Categories	Account	Nos. of followers
Government	@PMOIndia	4.24M
News Agency	@Reuters	5.82M
	@NDTVProfit	78.3K
English Business Dailies	@WSJ	5.58M
	@EconomicTimes	528K
	@bsindia	94.3K
Business Magazines	@forbes_india	76.7K
	@BT_India	20.7K
Financial Portals	@moneycontrolcom	61.5K
	@smartinvestor	12.5K

3.2 Methods and Material

For performing Sentiment Analysis, lexical based sentiment analysis has been chosen. **AFINN** lexicons. Inspired by **ANEW**, **Affective Norms for English Words** proposed by Bradley and Lang [22], Neilsen created **an AFINN list** [23] [24]. AFINN list is focused more towards web as it contains web-jargons, slang and obscene words. The list contains manually collected and scored English words rated for valence from minus five (negative) to plus five (positive). AINN-111 newest version which contains 2477 words and phrases has been used [25] [26]. Negative words are scored

from -1 to -5 similarly positive words are scored from 1 to 5. The range and better rating of the words is the reason why this lexicon is useful for sentiment strength analysis.

3.3 Text Processing

Twitter provides tweets of various qualities. It varies from high quality newswire to meaningless status updates. Abbreviations, unframed sentences, emoticons make the job of text processing tougher. Tweets from selected handlers of business magazines, financial portals, business news channels etc. were collected so the sentences were mostly well formed and mostly finance related news rather than general news.

3.4 Scoring Tweets

All the words in the AFINN- 111 list was divided into groups according to their scores.

1) **VeryNegativeTerms** -> words/phrases with scores -5 or -4 are classified under this heading.
2) **NegativeTerms** -> words/phrases with scores -3 or -2 or -1 are classified under this heading. eg. banned (-2), banish (1).
3) **PositiveTerms** ->words/phrases with scores 1 or 2 or 3 are classified under this heading.eg.cherished (2), helps (2), accept (1).
4) **VeryPostiveTerms** ->words/phrases with scores 4 or 5 are classified under this heading. eg. Hurrah (5), miracle (4).

For this study, sentiment has been classified as 'positive, negative or neutral' according to their score. Here are some examples:

Table 3. Sample of Tweet Orientation

Tweet	Score	Orientation
Hero acquires 60% stake in German bicycle company MIFA http://t.co/OOUGSwksHr	2	Positive
What is @Nissan_India's strategy for the road ahead? KenichiroYomura answers on #CEOonTheDrive tomorrow at 7:30 pm & Sun at 10:30 pm.	0	Neutral
#india #business : Global airline CEOs donâ€™t see SpiceJet troubles hurting Indian market: Global airline CEOs ...	-4	Negative

For finding the score (sentiment) of the tweets a parsing algorithm was designed. Tweets were read one at a time. First, the words were tokenized. Second, all the words were converted to lowercase and all the terms by separated using whitespace. Third, stops word, punctuations and URLs were removed. Then the tweets were ran against the Afinn List and by matching the words used in the tweets cumulative score of were calculated [27]. Lastly, LeveledScore of all the tweets was calculated by selecting the lowest score, assigning it a score of 0 and adjusting the score of other tweets accordingly.

3.5 Collecting Sensex and Oil & Gas Data

To find the correlation with different macroeconomic factors data for the same period from different authentic website of Sensex, NIFTY, Oil-Gas and Rupees-Dollar Exchange Rate was collected from various Government web sites

4. RESULT DISCUSSION

4.1 Correlation Analysis

Here we used **Pearson correlation**, defined as

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

Where value of r = 1 means prefect positive correlation and r = -1 means perfect negative correlation. If the value of r = 0.5 to 1 or -0.5 to -1, it indicates a high correlation, medium correlation if r = 0.3 to 0.5 or -0.3 to -0.5 else low correlation.

Next the correlation between the sentiment score and different macro-economic data collected over the same period of time was calculated.

Table 4. Correlation between Tweet Score and different Macro-Economic factors

Index	Period	Correlation
SENSEX	9 Months	0.670438
NIFTY	9 Months	0.677948
OIL-GAS	9 Months	0.422448
EXCHANGE RATE	9 Months	0.227962

The Sensex and Nifty indices are highly correlated with the tweet sentiment. There is a medium correlation of the crude oil with the tweet sentiment. But for the exchange rate it shows poor correlation.

The figure is below is week wise chart nifty and average tweet score, it has been calculated as a ratio.

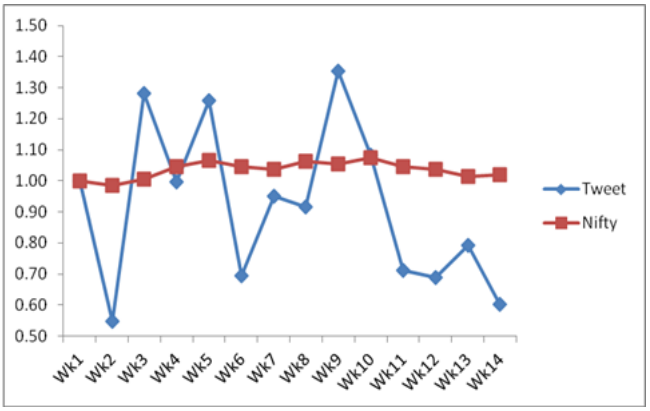

Fig 2: Nifty vs Average Tweet Score

It was observed that there was matching pattern between positive tweets and Nifty closing value as shown in fig 3.

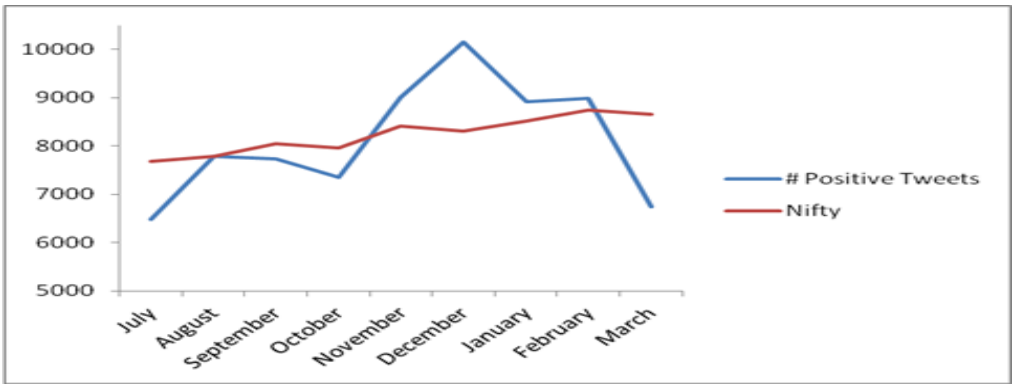

Fig 3: Comparison between #Positive Tweets and Nifty Closing Values

A matching pattern between daily tweet count and Nifty Turnover was observed which is shown in the fig 4 and Sensex closing value - Nifty closing Value - LeveledScore is shown in fig 5.

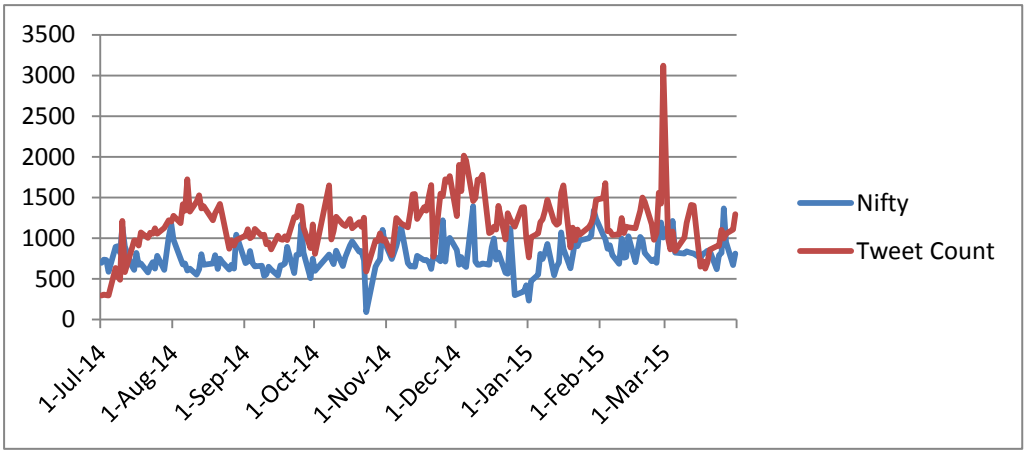

Fig 4: Matching pattern of Daily Tweet Count and Nifty Trunover

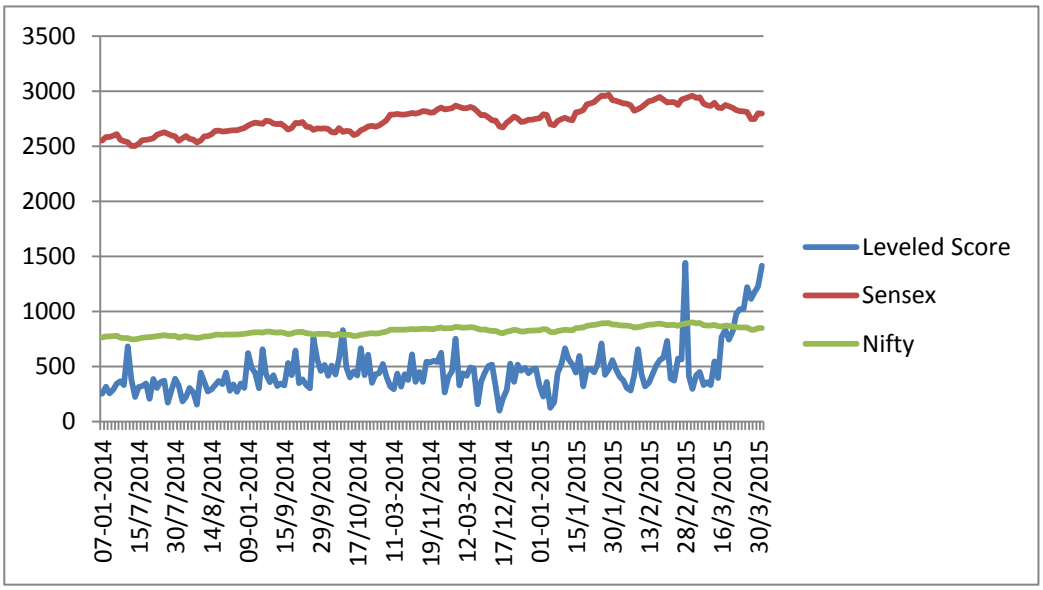

Fig 5: Sensex vs Nifty vs LeveledScore of Tweets

5. CONCLUSION AND FUTURE WORK

In this paper, an extensive study between sentiment from news and stock indices movement has been done. The relationship of sentiment, with other macro-economic factors has also been examined. For collecting news, 30 + twitter handles have been used. The study reveals good correlation with Sensex and Nifty and moderate correlation with exchange rate and Oil and Gas Prices.

In future, we intend to implement 'Bigrams' to get smooth instances in the case of negation features like 'not good' or 'not bad'. For better sentiment and mood analysis OpinionFinder and Google-Profile for Mood State (GPMOS) [28] which will give 6 dimensions to classify the tweets as happy, calm, alert, kind, sure and vitality. Stock market specific positive and negative words also need to be collected. NLP Techniques for identifying named entities can help to estimate stock specific sentiments.

6. REFERENCES

[1] Hawn, Carleen. "Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care." *Health affairs* 28.2 (2009): 361-368

[2] Sarasohn-Kahn, Jane. *The wisdom of patients: Health care meets online social media*. Oakland, CA: California HealthCare Foundation, 2008.

[3] Culnan, Mary J., Patrick J. McHugh, and Jesus I. Zubillaga. "How large US companies can use Twitter and other social media to gain business value." *MIS Quarterly Executive* 9.4 (2010): 243-259.

[4] Qualman, Erik. *Socialnomics: How social media transforms the way we live and do business*. John Wiley & Sons, 2010.

[5] Fama, E. F. (1965) The Journal of Business 38 , 34–105

[6] Jensen, Michael C. "Some anomalous evidence regarding market efficiency." *Journal of financial economics* 6.2 (1978): 95-101.

[7] Bollen, Johan, Huina Mao, and XiaojunZeng. "Twitter mood predicts the stock market." *Journal of Computational Science* 2.1 (2011): 1-8.

[8] Zhang, Xue, HaukeFuehres, and Peter A. Gloor. "Predicting stock market indicators through twitter "I hope it is not as bad as I fear"." *Procedia-Social and Behavioral Sciences* 26 (2011): 55-62.

[9] Morstatter, Fred, et al. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." (2013).

[10] Chung, Jessica Elan, and EniMustafaraj. "Can collective sentiment expressed on twitter predict political elections?." *AAAI*. 2011.

[11] Tumasjan, Andranik, et al. "Election forecasts with Twitter: How 140 characters reflect the political landscape." *Social Science Computer Review* (2010): 0894439310386557.

[12] Tumasjan, Andranik, et al. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10 (2010): 178-185.

[13] Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., &Flammini, A. (2011, July). Political polarization on twitter.In *ICWSM*.

[14] Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1, pp. 492-499. IEEE, 2010.

[15] Bollen, Johan, Huina Mao, and XiaojunZeng. "Twitter mood predicts the stock market." *Journal of Computational Science* 2.1 (2011): 1-8.

[16] Zhang, Xue, HaukeFuehres, and Peter A. Gloor. "Predicting stock market indicators through twitter "I hope it is not as bad as I fear"." *Procedia-Social and Behavioral Sciences* 26 (2011): 55-62.

[17] Chang, Hsia-Ching. "A new perspective on Twitter hash tag use: diffusion of innovation theory." *Proceedings of the American Society for Information Science and Technology* 47.1 (2010): 1-4.

[18] Cullum, Brannon. "What makes a Twitter hashtag successful." *Movements.org*17 (2010).

[19] Blaszka, Matthew, Lauren M. Burch, Evan L. Frederick, Galen Clavio, Patrick Walsh, and J. Sanderson. "# WorldSeries: An empirical examination of a Twitter hashtag during a major sporting event." *International Journal of Sport Communication* 5, no. 4 (2012): 435-453.

[20] Chang, Hsia-Ching, and HemalataIyer. "Trends in Twitter hash tag applications: Design features for value-added dimensions to future library catalogues."*Library Trends* 61, no. 1 (2012): 248-258.

[21] Davidov, Dmitry, Oren Tsur, and Ari Rappoport. "Enhanced sentiment learning using twitter hashtags and smileys." *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010.

[22] Bradley, M. M., and Lang, P. J. Affective Norms for EnglishWords (ANEW) Instruction Manual and Affective Ratings.Technical Report C-1, The Center for Research inPsychophysiology University of Florida, 2009.

[23] Nielsen, Finn Årup. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs." *arXiv preprint arXiv:1103.2903* (2011).

[24] Hansen, Lars Kai, et al. "Good friends, bad news-affect and virality in twitter."*Future information technology*.Springer Berlin Heidelberg, 2011.34-43.

[25] Pavlopoulos, Ioannis, and Ιωάννης Παυλόπουλος. "Aspect based sentiment analysis." (2014).

[26] Stuefer, Marina. "Social Media Sentiment Analysis for Stock Price Behavior Prediction."

[27] Zhang, Xue, HaukeFuehres, and Peter A. Gloor. "Predicting stock market indicators through twitter "I hope it is not as bad as I fear"." *Procedia-Social and Behavioral Sciences* 26 (2011): 55-62.

[28] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." *Proceedings of the Workshop on Languages in Social Media*.Association for Computational Linguistics, 2011.