# Comparative Analysis of the Impact of Blogging and Micro-blogging on Market Performance

Devendra Tayal
Department of Computer Science and Engineering
Indira Gandhi Institute of Technology
New Delhi, India
(dev_tayal2001@yahoo.com)

Satya Komaragiri
Associate Software Engineer
Red Hat Software Services India Pvt. Ltd.
Pune, India
(satya.komaragiri@gmail.com)

*Abstract* — **The general perceptions about a product and the reputation of the company determine to a great extent how well the product sells. It is thus imperative that we make efforts to understand the public opinions and sentiments, as they can be a very good indicator of the product's future sales performance.**

**In this paper, we explore the two most common online media which have been used by the public to express such type of subjective content: Blogs and Micro-blogs. We perform a comparative analysis of the predictive power of the two media to know which of these formats can prove to be a more useful representative of sentiments to an autonomous stock price prediction system.**

***Keywords – Blogging; Micro-blogging; Sentiment Analysis; Text Mining; Opinion Mining***

## I. INTRODUCTION

For many businesses, online customer opinions have become a type of virtual currency that can make or break their products. With sentiment analysis algorithms, companies can identify and assess the wide variety of opinions found online and create computational models of human opinion [1].

Weblogs (blogs) are often online diaries published in reverse chronological order, and they can also be commentaries or discussions on a particular subject, ranging from mainstream topics (e.g., food, music, products, politics, etc.), to highly personal interests [2]. Blogging has grown by 68% in 2008 and 77% of active web users read blogs [3].

A lot of research based effort has been put in analyzing the blogosphere. [4-6] are just a few examples of the efforts made in this field. These works ascertain beyond doubt that the sentiments contained in blogs do represent the public opinion and that there is a correlation between the sentiments expressed in the blogs and the performance of the stocks. With more and more enhancements to the sentiment analysis algorithms and text mining techniques, the correlation will only increase making blogs an indispensable medium of expression that the companies must pay heed to.

With the immense success of blogs as a means to present subjective content like opinions and sentiments, newer means of expressing feelings like micro-blogging (Twitter etc.) have emerged. Providing a unique blend of social networking ideas with blogging, micro-blogging is gaining widespread acceptance as the new tech-savvy way of sharing information.

According to market researcher Nielson Online the number of unique visitors to Twitter[1] jumped an unbelievable 1,382 percent, from 475,000 in February 2008 to 7 million in February 2009 [7]. The recent Twitter campaign Moonfruit[2] Inc. threw up some very interesting statistics. The 7 day campaign increased the traffic to Moonfruit.com by 600% and the number of signups by 100% [8].

With an ever-increasing user base, it is perhaps prudent to investigate the extent to which it can represent the general public sentiments.

For this purpose, we build a foreteller system which predicts the future performance of a stock by using sentiment analysis on relevant tweets (twitter posts) to study the predictive power of micro-blogging. We perform a case study on two well-known technological companies with recent product launches. We pit the results thus obtained against those obtained by using the same software on relevant blogs as the source of sentiment containing text. We then summarize the results to see how well micro-blogging performs as compared to blogging.

## II. RELATED WORK

The current work on micro-blogging, more specifically Twitter, concentrates on studying whether or not it is a phenomenon capable of making a difference, what drives it and what could be its other possible uses and future.

Freiert [9] studied the demographics of Twitter users and categorized them by age, gender and location. Java *et. al.* [10] classified Twitter users as *information sources*, *information seekers* and *friends.* They also classify

[1] http://twitter.com/

[2] http://www.moonfruit.com/macbook-pro.html



ISSN : 0975-3397

intentions as *daily chatter, information sharing, reporting news and conversation.* Studies by HP lab [11] prove that it is not the easily visible follower-followed network which is the driving force behind twitter but the more subtle network of 'friends'- a term they define as a link between two people who have sent at least 2 messages to each other. Honeycutt and Herring [12] explore Twitter to find how much potential it holds for collaboration.

In our paper, we explore the potential of micro-blogging in terms of being able to predict the future stock values. For this purpose, we first take a look at how the information on the Blogosphere has been used to extract such information.

There are currently two main approaches to analyzing the Blogosphere. One approach is to make use of links or URLs in Blogosphere [4]. A time graph for Blogosphere can be built and be viewed as a function of time. This graph can then be studied to analyze the changing trends. This method takes the number of blog mentions and the internetwork of the blogs into account but does not care for the sentiment content of the blogs.

The second method addresses the aforementioned issue. It concentrates primarily on the sentiments as expressed in the blogs viz. legal blogs [13], movie reviews [4, 6], product reviews [14, 15], e-learning [16] or even profile generation based on [17] watching habits. The impact of different types of blogs has been studied thoroughly via sentiment analysis.

Verlic *et.al.* [18] present a very interesting study where they employ sentiment analysis on the scientific papers published at the CBMS symposiums from a content analysis point of view to see if there are any significant differences in psycho-social texture of the accepted papers.

Sentiment analysis takes linguistics into consideration. It is more complicated as subjective feelings are hard to be represented by a computational model as opposed to statistics used in previous methods.

Grzegorz *et. al.* [6] combine 3 methods: linguistic classifier, group behavior classifier and statistic classifier to classify opinions on movie reviews and combine the results to make the final assessment.

Kanayama *et. al.* [19] show how machine translation is similar to sentiment analysis and propose a high-precision sentiment analysis system at a low development cost, by making use of an existing transfer-based machine translation engine.

Sentiment analysis algorithms be used to assign positive and negative polarities to the opinions at the document level or can go deeper down to sentence, subject and phrase level [15, 20]. Some methods at the parts of speech like looking for the subject and the object [5]. Our system narrows down the scope of investigation to sentence level. As the number of tweets posted every second is much more than the number of blogs posted, this method puts both media on equal footing as the number of sentences in a blog is greater.

These Sentiment analysis algorithms are either semantic based or learning based. These approaches concentrate on opinion word collection in the form of a sentiment directory or a large scale knowledge base to assign sentiments. The other learning-based approaches for sentiment classification typically leverage the manually labeled documents as the training set, and then apply traditional learning techniques, such as, Naïve Bayes, Maximum Entropy and SVMs, to do sentiment classification [21]. In our work, we choose the former approach as we are interested in making a generic tool which can be used to study any new form of media and we feel it is rather infeasible to train a system with sufficient number of known samples in the case of newer media.

The concept of positive and negative polarity of sentiments was extended to include the degree of positivity and negativity on the scale of -1 to 1 by Subrahmanian & Reforgiat [22]. In their work, they propose an 'Adjectives, Verbs and Adverbs' (AVA) framework which extends Beth Levin's verb classification to incorporate sentiment information. We choose a scale of -5 to 5 to rate our sentiment terms.

In their work, Yang Liu *et. al.* [4] propose the Auto Regressive Sentiment Aware (ARSA) model for product sales prediction, which reflects the effects of both sentiments and past sales performance on future sales performance. In our work, we take both the factors into consideration although we differ in the method chosen. We use the sentiment scores calculated as the measure of the estimated difference (referred to as $\Delta$ here on) in the stock prices while keeping the previous stock values as the centre point.

We employ Whitelaw *et. al.*'s [23] 'Adjectival Appraisal Groups' for constructing our sentiment repository. We make some modifications to the method used by introducing the rating factor which differentiates the degree of positivity and negativity based on the adjective (fabulous conveys a higher degree of positivity than good) in addition to the presence of phrases like 'every' or 'not so' used by Whitelaw et. al.

In the present paper, our aim is to study the predictive power of micro-blogging as compared to blogging. The sentiment analysis algorithm used is a tool to that end. We have made some improvements in the sentiment analysis algorithm as described above and in more detail in section III-C. We not only classify the posts as positive or negative but also quantify the positivity and negativity to be able to use it for prediction.

## III. DESCRIPTION OF OUR SYSTEM

### *A. Overall Description*

To carry out our study, we implemented a system in



ISSN : 0975-3397

Java which takes a stock name as input, downloads the relevant blog and micro-blog information from the Internet and performs certain pre-processing activities to obtain pure text. This process is described in detail in section III-B. It then performs sentiment analysis firstly on the blog text and then on the micro-blog text using a sentiment repository. The sentiment repository is created using semi-automated techniques as explained in section III-C. The results of sentiment analysis and previous stock price history of that stock are then used to make two predictions for the next day's stock value. One, using the sentiment analysis results obtained from blog contents and second using results obtained from micro-blog contents. The results are logged in the database module (section III-E).

The above procedure is repeated several times on the same stock to fine-tune our system with that stock (Section III-D. This can be referred to as the training period. The longer is the training (i.e. the more the runs on the same stock), the more accurate the predictions become, as the prediction module is better able to quantitatively correlate the sentiment score to the fluctuation in stock prices for that stock with each new run. The full system is also implemented in Java.

The following subsections explain the individual tasks of our system along with their implementation details.

### *B. Web Interface and Data Preparation*

1) *Obtaining Blog Contents:* To obtain the blog contents, we implement a system in Java which semantically crawls the web to collect the relevant blog links. The starting link is constructed by examining the URLs generated by a popular blog search engine[3]. The content from the links are then downloaded automatically by the system. The data from those links are then scraped to obtain the blog contents on which we can run our algorithm. The contents are cleaned further using regular expressions to obtain pure text.

2) *Obtaining Micro-blog Contents:* For this purpose we use Twitter, the most popular micro-blogging service, as a general representative of the kind of sentiment statements that can be found on micro-blogging sites. The significantly high number of Tweets posted per minute can give us the same amount of sentiment content as the blogs we consider.

To obtain the contents from micro-blogs, the system used for blogs was adapted to use the search tool which is a part of the Twitter API[4] instead of the web crawler. The concept of hash tags[5] makes it even more convenient. The scraper was also modified to extract the Tweets. As in the case of blogs, we use regular expressions to clean out the contents to extract pure text.

### *C. Sentiment Analysis*

Once the text has been obtained, the sentiment analysis is performed. To this end, we use the concept of adjectival appraisal groups. Appraisal groups are defined as those groups and phrases in a text that tells about the kind and intensity of appraisal expressed.

We created a sentiment repository by following the procedure used by Whitelaw *et. al.* [23] to build their lexicon. A semi-automated technique is used to quickly build up a repository using the seed terms given for various appraisal options in [24] and [25]. Modifier seed terms were generated similarly, by finding adverbs collocating with adjective seed terms in our corpus. Candidate expansions for each seed term were generated from WordNet and from two online thesauri[6] as described in [23].

Each ranked list was manually inspected to produce the final set of terms used. We enhanced the method by incorporating an "adjective rating" which accounts for the relative degree of positivity or negativity amongst adjectives. We use a 10 point scale of -5 to 5 where -5 is maximally negative and +5 is maximally positive. The modifiers were represented as formulae which would be applied on the base adjective's rating to obtain a sentiment score for the entire appraisal group. To account for the micro-blogging lingo we added a provision to read *#sentiment_term* as *sentiment_term.* These accounts for posts which very clearly and succinctly express their sentiments in terms like #fail.

The above modifications have a two-fold benefit. Firstly, it helps us represent more accurately the degrees of intensity that the adjectives hold inherently within them, irrespective of the presence of modifiers. Secondly, our system leads to numeric values being calculated for all appraisal groups as they are encountered. This is a very important requirement as we need to find a way to quantify sentiments in order to use them for automatic prediction.

We use the repository thus created to perform sentiment analysis on the blog and micro-blog contents one by one. We used an implementation of Brill's [26] part-of-speech tagger to help us find adjectives and modifiers. For every appraisal group found, the sentiment score is added to the total score so far for that medium. The final score for the medium is then normalized by dividing by the number of sentences parsed for that medium. This is done in order to minimize the effect of unequal amount of text downloaded for the two media.

### *D. Prediction of Stock Price*

As mentioned in section I, we use both, the previous stock values and sentiment analysis together for predicting

[3] http://blogsearch.google.com/

[4] http://search.twitter.com/

[5] http://hashtags.org/

[6] http://m-w.com and http://thesaurus.com



ISSN : 0975-3397

the future stock values. The prediction is done for both the media separately using the sentiment score and history for the medium under consideration.

We do this by using the present day's stock value as the base and adding a Δ (can be either positive or negative) to it (where Δ is a function of the current final sentiment score of the medium) and the previous history for this medium of all runs of the software on the same stock name. The function becomes more and more accurate with each run and thus the Δ is fine-tuned to the particular stock. This results in increasing accuracy of prediction with each run on that stock.

Once the final score is obtained for both the media, they are stored in the database along with the history as the next day's predicted stock value for later verification. This entry also serves as a part of the history used for calculation of Δ for the next run.

### *E. Database*

The database stores the history for all the stocks on which the system has been run to be used for learning of the system and also to store the results obtained for later analysis. The entry for each stock contains the log of all dates the system was run, the then-present stock value, the value predicted using blog entries and the value predicted using micro-blog entries.

## IV. WORKING METHODOLGY

The System was continuously run over a period of three months. Firstly we trained[7] the system for a month on two well known technology companies - Google Inc. and Microsoft Corp., as they are amongst the most talked about companies in the blogging and micro-blogging circles.

We then used our trained model to predict the stock values for the next two months. During this period, Microsoft (the operating system giant) and Google (the web search giant) were both in a competing phase. It was noted that Google announced Wave at the same day as Microsoft launched Bing. Moreover, Google also announced Google Chrome OS thus inviting many blogs and micro-blogs posts which were very high in their sentiment content. This provided a great opportunity to us to test our product.

## V. RESULTS AND INTERPRETATION

The stock values and the predicted values calculated using both media pertaining to Google Inc. are shown below:

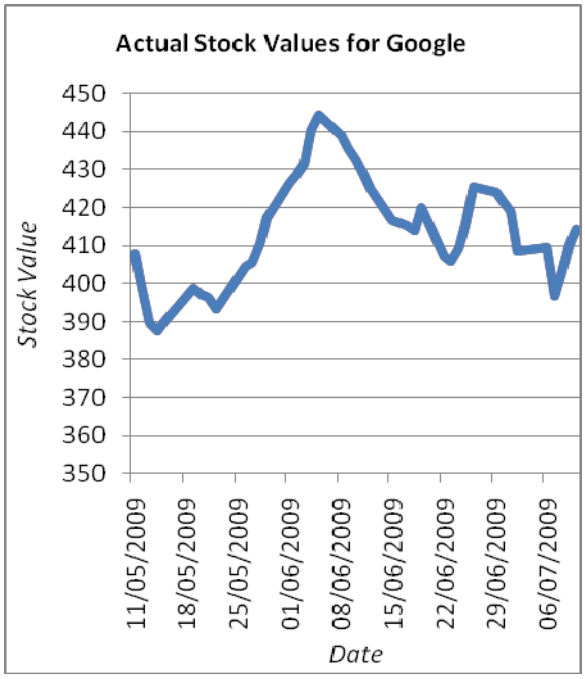


Figure 1. Actual stock values for Google.

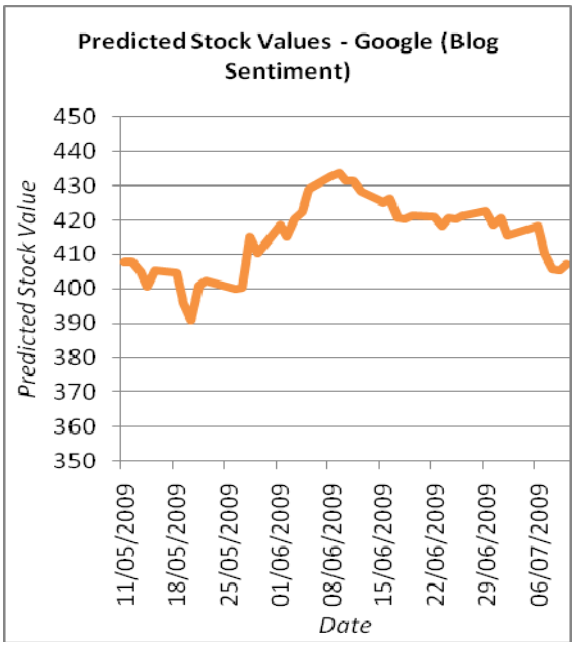


Figure 2. Predicted stock values of Google using blog sentiments.

[7] It may be noted again that by training we do not mean the classical training in which some sample corpus is manually rated and the system learns to rate unknown corpus by using machine learning algorithms. We refer to training as described in section III-A.



ISSN : 0975-3397

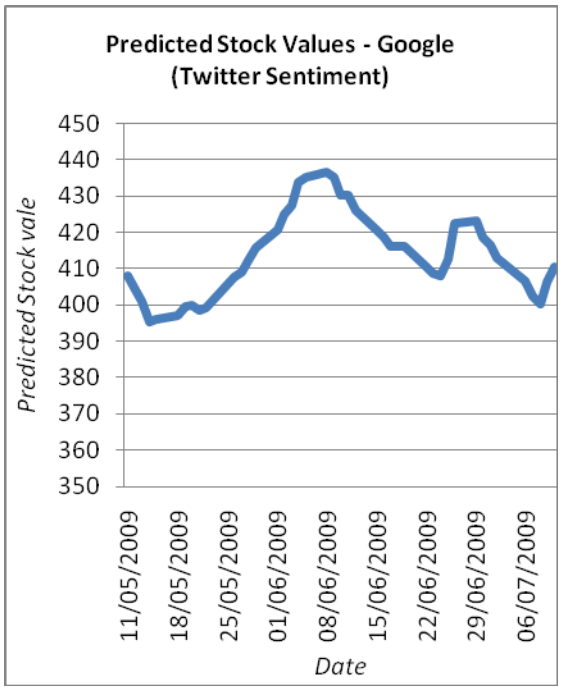


Figure 3. Predicted stock values of Google using micro-blog sentiments

The experimental results for Microsoft Corporation are as shown:

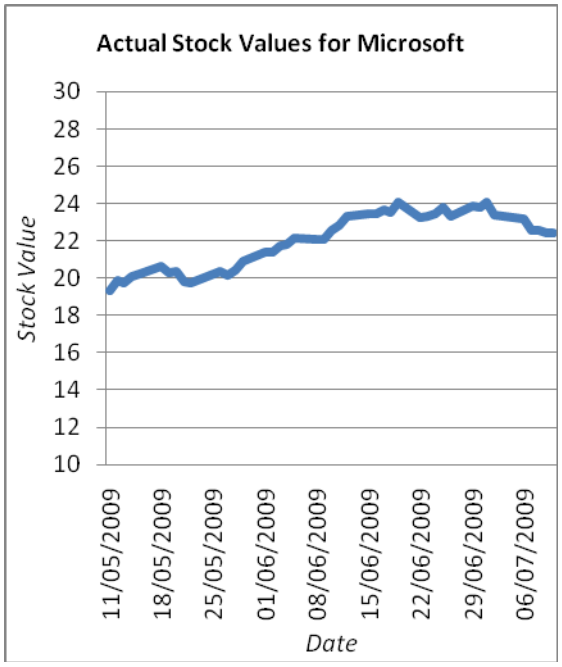


Figure 4. Actual stock values of Microsoft.

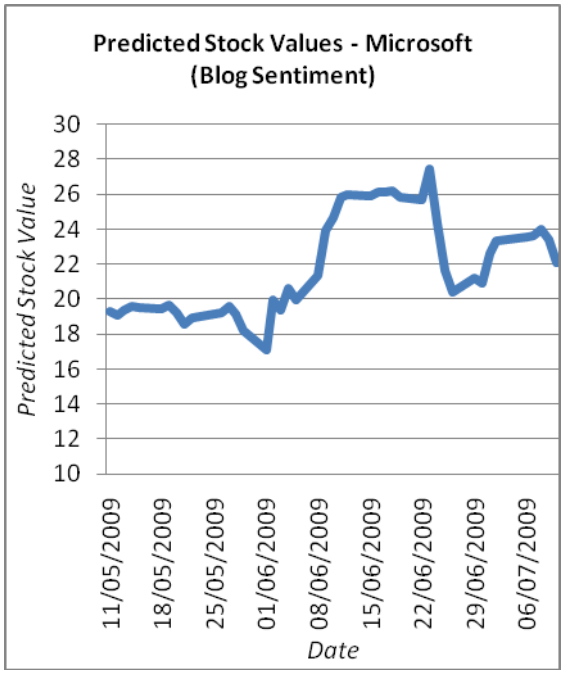


Figure 5. Predicted stock values of Microsoft using blog sentiments.

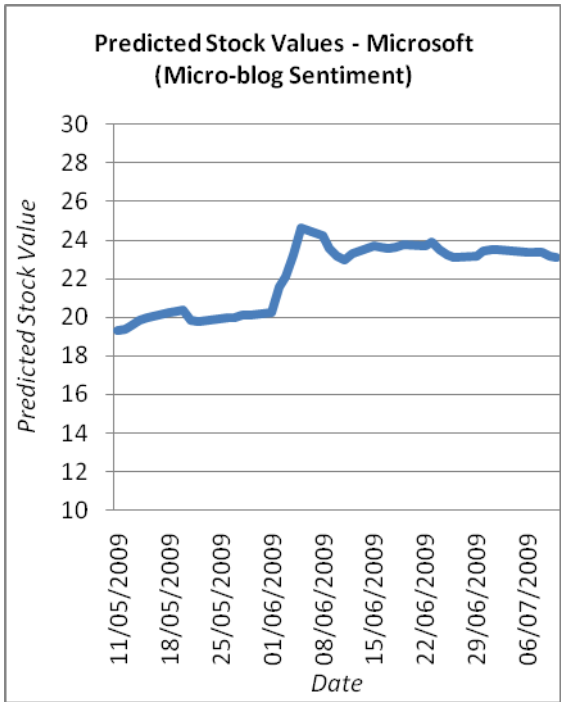


Figure 6. Predicted stock values of Microsoft using micro-blog sentiments.

In Google's case, the correlation between the actual stock prices and the predicted prices using blog sentiments was found to be 0.811 whereas the correlation between the actual stock prices and the predicted prices using micro-blog sentiments was found to be a staggering 0.972

For Microsoft, correlation between the actual stock



ISSN : 0975-3397

prices and the predicted prices using blog sentiments in this case was found to be 0.766 whereas the correlation between the actual stock prices and the predicted prices using micro-blog sentiments was noted to be 0.911.

The above results clearly show that analyzing micro-blogs is a more reliable means to predict the future performance of a company than blogs.

Many factors could have contributed to this outcome. For one, sentiment analysis on micro-blogs is more accurate as the text is limited by the nature of the medium. A limit of 146 characters means that the posters have to be more concise. This reduces the scope for elaborate circumlocution and use of phrases that are hard to be fit into a recursive algorithm. In other words, the written text is more representative of the actual intent of the poster.

Moreover, the current state of Natural Language Processing is not developed enough to judge which portions of the text are unrelated accurately. As multiple topics cannot be clubbed into a single post (or a tweet) in case of micro-blogs, this problem of removing unrelated content (which might be rich enough in sentiment words to mislead our algorithm) is not as severe as in the case of blogs.

Another less obvious but important factor could be the social networking nature of this medium. With concepts like re-tweets that help broadcast the ideas to reach a bigger audience and reply-tos that trigger conversation chains in which all replies, as they are limited to 146 characters too, provide for more concise sentiment data for us.

## VI. CONCLUSION

In our work, we sought to investigate the potential micro-blogging holds for stock prediction. As the importance of the role of the sentiments posted on blogs has been well established, we perform a comparative study of the predictive power of blogs and micro-blogs. We find that the correlation between the values predicted by using micro-blog sentiments is higher than the correlation obtained by using the blog sentiments. In our experiments, micro-blogs consistently outperformed blogs in their predictive capacity. It is hence clear that micro-blogs can be a useful means for predicting the future performance of stocks. This is an important finding from the perspective of both, the companies and the stockholders.

## VII. REFERENCES


[1] Alex Wright, '*Our sentiments, exactly*' Communications of the ACM Volume 52 , Issue 4 (April 2009) A Direct Path to Dependable Software, COLUMN: News, Pages 14-15, Year of Publication: 2009 ISSN:0001-0782

[2] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins, *Structure and Evolution of blogspace*. Commun. *ACM*, 47(12):35–39, 2004.

[3] Libby Varcoe (2009, Feb 23) *Blogging grows by 68%.* Retrieved July 13, 2009 from libbyvarcoe.wordpress.com: http://libbyvarcoe.wordpress.com/2009/02/23/blogging-grows-by-68/

[4] Y. LIU, X. HUANG, A. AN, *et. al.* (2007) *ARSA: A Sentiment-Aware Model for Predicting Sales Performance using Blogs.* Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. Amsterdam, The Netherlands. ACM: 607-614.

[5] Navendu Garg, Kenneth Bloom, Shlomo Argamon, (2006) '*Appraisal Navigator*', Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. Seattle, Washington, USA

[6] Grzegorz Dziczkowski, Katarzyna Wegrzyn-Wolska, *'An Autonomous System Designed for Automatic Detection and Rating of Film Reviews'* wi-iat, vol. 1, pp.847-850, 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008

[7] Jin L. (2009, June 21). *Businesses using Twitter, Facebook to market goods.* Retrieved July 13, 2009 from post-gazetter.com: http://www.post-gazette.com/pg/09172/978727-96.stm

[8] Kunal Kripalani (2009, Jul 7)*Moonfruit Twitter Campaign Analysis.* Retrieved July 13, 2009 from Social Media Guidelines http://www.social-bug.com/moonfruit-twitter/

[9] Freiert, M. (2008, May 15). *Twitter traffic explosion: Who's behind it all?* Compete. Retrieved May 21, 2008 from http://blog.compete.com/2008/05/15/twitter-traffic-growthusage-demographics/

[10] Java, A., Song, X., Finn, T., & Tseng, B. (2006, August).*Why we Twitter: Understanding micro-blogging usage and communities.* Joint 9th WEBKDD and 1st SNA-KDD Workshop e07, San Jose, CA

[11] Bernardo A. Huberman, Daniel M. Romero and Fang Wu Social Computing Laboratory, HP Labs(2009. January). ''*Social networks that matter: Twitter under the microscope*. First Monday, Vol. 14 http://www.hpl.hp.com/research/scl/papers/twitter/twitter.pdf

[12] Courtenay Honeycutt, Susan C. Herring. '*Beyond Micro-blogging*: *Conversation and Collaboration via Twitter'.* In proceedings of the 42nd Hawaii International Conference on System Sciences. Pages 1-10. Year of Publication: 2009, ISBN:978-0-7695-3450-3

[13] Jack G. Conrad , Frank Schilder, '*Opinion mining in legal blogs'*, Proceedings of the 11th international conference on Artificial intelligence and law, June 04-08, 2007, Stanford, California

[14] Zhu Zhang, '*Weighing Stars: Aggregating Online Product Reviews for Intelligent E-commerce Applications*' Intelligent Systems, IEEE , vol.23, no.5, pp.42-49, Sept.-Oct. 2008

[15] Yi, J.; Nasukawa, T.; Bunescu, R.; Niblack, W., '*Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques*'. ICDM 2003. Third IEEE International Conference on Data Mining, , vol., no., pp. 427-434, 19-22 Nov. 2003

[16] Dan Song; Hongfei Lin; Zhihao Yang '*Opinion Mining in e-learning System*' NPC Workshops. IFIP International Conference on Network and Parallel Computing vol., no., pp.788-792,18-21 Sept. 2007

[17] Yasufumi Takama, Yuki Muto, '*Profile Generation from TV Watching Behavior Using Sentiment Analysis*' wi-iatw, pp.191-194, 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, 2007

[18] Verlic, M.; Stiglic, G.; Kocbek, S.; Kokol, P., '*Sentiment in Science - A Case Study of CBMS Contributions in Years 2003 to 2007*' Computer-Based Medical Systems, 2008. CBMS '08. 21st IEEE International Symposium on , vol., no., pp.138-143, 17-19 June 2008

[19] Kanayama Hiroshi , Nasukawa Tetsuya , Watanabe Hideo, '*Deeper sentiment analysis using machine translation technology'*, Proceedings of the 20th international conference on Computational Linguistics, p.494-es, August 23-27, 2004, Geneva, Switzerland

[20] Tetsuya Nasukawa , Jeonghee Yi, '*Sentiment analysis: capturing favorability using natural language processing'*, Proceedings of the 2nd international conference on Knowledge capture, October 23-25, 2003, Sanibel Island, FL, USA




ISSN : 0975-3397

[21] Keke Cai, Scott Spangler, Ying Chen, Li Zhang, '*Leveraging Sentiment Analysis for Topic Detection*', IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008 vol. 1, pp.265-271, 2008

[22] Subrahmanian, V.S.; Reforgiato, Diego, '*AVA: Adjective-Verb-AdverbCombinations for Sentiment Analysis',* Intelligent Systems, IEEE ,vol.23, no.4, pp.43-50, July-Aug. 2008

[23] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. '*Using appraisal groups for sentiment classification'.* In Proc. Conference on Information and Knowledge Management, Bremen, Germany, 2005.

[24] J. R. Martin and P. R. R. White. '*The Language of Evaluation: Appraisal in English'*. Palgrave, London, 2005. (http://grammatics.com/appraisal/).

[25] Christian Matthiessen. '*Lexico-grammatical cartography: English systems'.* International Language Sciences Publishers, 1995.

[26] Eric Brill. '*A simple rule-based part of speech tagger'.* In proceedings of ACL Conference on Applied Natural Language Processing. Trento, Italy. 1992.

## VIII. ACKNOWLEDGEMENTS

The authors express their gratitude to Indira Gandhi Institute of Technology, GGSIP University, Delhi and Red Hat Software Services India Pvt. Ltd. for providing the resources to carry out the present research work. The authors also thank Sun Microsystems India for seeing the potential and rewarding the initial attempt to quantify sentiments in blogs and using them to predict stock performance.

## IX. AUTHORS' PROFILES

Dr. Devendra Kumar Tayal was born in 1977 in Delhi, India. He has acquired the degrees of B.Sc. (H) (Mathematics), M.Sc. (Mathematics), M.Tech. (Computer Engineering) & Ph.D. (Computer Engineering) from Jawaharlal Nehru University, Delhi, India. He is currently serving as Associate Professor and Head, in Department of Computer Science & Engineering, Indira Gandhi Institute of Technology, GGSIP University, Delhi. He has written about a dozen Research papers in International Journals and an equal number in International Conferences. He is also a member of International Advisory Board on International Journal of Software Engineering & its Applications, Korea and International Journal of Computer Science, Hongkong. His research areas include databases, intelligent systems, data mining and software Engineering.

Satya Komaragiri was born in 1987 in India. She has done her B.Tech. in Computer Science & Engineering from Indira Gandhi Institute of Technology, Delhi and is currently working as an associate software engineer at Red Hat Software Services India Pvt. Ltd. Her research areas include operating systems, algorithms, speech recognition and natural language processing.



ISSN : 0975-3397