ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 3, Issue 1 (Jan. - Mar. 2015)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2015)

A Model of Opinion Mining for Classifying Movies

[I]Sohom Ghosh, [II]Santanu Modak, [III]Abhoy Chand Mondal
[I]Dept. of CSE, Heritage Institute of Technology, Kolkata, West Bengal, India
[II,III]Dept. of Computer Science, University of Burdwan, West Bengal, India.

Abstract
Opinion analysis has become a flourishing frontier as of late. In this paper, we exhaustively study movie reviews from a popular online database. We randomly sample more than 1000 reviews with titles to train our model. It is capable of suggesting words during the process of appraising a film. It can intelligently anticipate the words that an appraiser is going to use from the title of his opinion. Furthermore, it has the potential to learn. Whenever it finds that it is unable to suggest words, it learns from the critic's opinion. Moreover, this innovative model is able to compute the popularity of a film by examining the opinions. Thus, it simplifies the job of reviewing by making it quicker and effective. It labels a movie as 'super-flop', 'flop', 'cool', 'hit' or 'super-hit' based on what the reviewers opine.

Keywords
Opinion Mining, Sentiment Analysis, Natural Language Processing, Polarity Computation, Recommendation System, Machine Learning, Collaborative Filtering.

I. Introduction

Opinion Mining is a recent research area in the Natural Language Processing community. Since the Research in Opinion Mining is not concerned with the topic of document, but the detection of actual sentiment it express. Within Opinion Mining, several subtasks can be identified, like, detection of sentiment, calculation of polarity, classification of sentiment, authorship identification, spam opinion detection, opinion summarization etc.

In this paper we confront the task to classify movies with labels as 'super-flop', 'flop', 'cool', 'hit' or 'super-hit' based on user's opinion towards the movie. To classify opinions, most of the existing approaches rely on training from human-annotated data. But we propose a different approach in this paper.

The remaining part of the paper is organized as follows. In Section 2 we review the related works in opinion mining. Our problem definition is stated in details in section 3. Section 4 explains the terms and notations which are used in this paper. Section 5 describes the datasets gathered from popular movie review website. In Section 6 we present a flow chart for suggesting word during reviewing and discovering opinion polarity. The results of conducted experiments are reported in Section 7. In Section 8 we present a comparative study with some popular movie review websites. Section 9 deals with methods of fine tuning the model. We conclude the paper in section 10 mentioning its future work.

II. Literature Review

Much progress has been done in the field of opinion mining in the last few years. Several supervised and unsupervised approaches are proposed. Sima and Vunvulea proposed a rule-based opinion mining and holder identification technique for cross domain analysis [1]. Rustamov et.al. proposed a sentence level subjectivity analysis based on occurrence of word in the corpus using Hidden Markov Model. This approach does not need any linguistic knowledge, so it can be applied to any language [2]. Jusoh and Alfawareh proposed fuzzy lexicon based approach to determining degree of positivity or negativity [3]. Wang and Zhou applied MRA (mutual reinforcement approach) to improve the accuracy of the mining results [4]. Kamal and Abulaish proposed a sentiment analysis system which combines rule-based and machine learning approaches to identify feature-opinion pairs and their polarity [5]. Rustamov, Mustafayev and Clements attempted to detect sentence-level subjectivity by means of hidden Markov model which hasn't been thoroughly investigated for subjectivity analysis and also proposed a feature extraction algorithm which calculates a feature vector based on the statistical occurrences of words in a corpus without any linguistic knowledge except tokenization [6].

III. Problem Definition

Given the name of a movie and title of an opinion about it, we predict the words an appraiser will probably use. We anticipate the score from what he opines. We have built our unique word corpora separate for positive and negative words. We compare the adjectives and adverbs of the opinion with our corpora to detect the degree of polarity of the film. We make a comparative study of our model to that of ID's and RT's. Finally, we use Artificial Neural Networks (ANN) to fine-tune our model.

Note: ID, RT are two popular online movie reviewing websites.

IV. Terms and Notations

This section defines some general terms which are used throughout the paper and relevant to our work.

A. Tokenization

The process of splitting a sentence into its constituent words is known as Tokenization. NLTK (Natural Language Toolkit) provides us with Punkt sentence tokenizer. Our model uses split() function to tokenize sentences. After removing trailing blank spaces using strip(), we are splitting the corpus whenever we come across white spaces. Example: Original Sentence: - "The Sky is Blue"; After Tokenization: ['The','sky','is','blue']

B. POS Tagger

POS tagger (or parts of speech tagger) is an inbuilt package of NLTK which maps each word to the parts of speech they belong. Example: Original Sentence: "The opening aerial shots of the prison are a total eye-opener." After using POS tagger: - (S The/NNP opening/VBG aerial/JJ shots/NNS of/IN the/DT prison/NN are/VBP a/DT total/JJ eye-opener. /NNP).

C. Stemming

Words like 'behaving', 'behave', 'behaved' means the same. They have structural affixes which changes there spelling. To simplify our task we convert these words to a single form i.e. is 'behave'.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

This act is referred to as stemming. We use Porter stemmer as a pre-processing step.

D. Corpus

Corpus means collection of words. We have built unique corpora separate for positive and negative words. It contains around 6000 words. So, let's look at a sample from it.
Positive words: - [':)', 'absolutely', 'bounty', 'calm', 'meritorious', 'skilled', 'wow', 'zeal']
Negative words: - [':(', 'abysmal', 'hurtful', 'ignore', 'malicious', 'worthless', 'yucky']

E. Polarity

Polarity refers to the positivity or negativity of a word. In this paper, we label positive words as '+1' and negative words as '-1'.

F. Mapping

Mapping is representation of relations. It refers to a function. For example: *f: X-> Y,* denotes that *f* is a function which maps X to Y. In this paper we use one-to-many and many-to-one mapping.

G. Collaborative Filtering

For building the recommendation system to suggest words, we use collaborative filtering. It refers to the art of proposing words by gathering interests from the users (collaboration).

H. Mean

Mean is a statistical term. It is also referred to as the average. Here, we are finding out the arithmetic mean of some discrete values. The formula for mean is:-

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i \qquad \ldots (1)$$

Here, A denotes the arithmetic mean, a_i denotes the discrete values, Σ represents summation and n refers to the number of discrete values.

I. ID's formula to calculate scores for top 250 movies

$$W = \frac{Rv+Cm}{v+m} \qquad \ldots (2)$$

Here, *W*= weighted rating, *R*= average for the movie as a number from 0 to 10 (mean) = (Rating), *v*= number of votes for the movie = (votes), *m*= minimum votes required to be listed in the Top 250 (currently 25,000), *C*= the mean vote across the whole report (currently 7.0)

J. RT way of labelling movies

As per RT method of labelling, a movie is tagged as 'Fresh' if more than 60% of its reviews are positive. Otherwise, it is tagged as 'Rotten'. 'Audience rating' refers to the % of users who rated the movie above 3.5 out of 5.
Note:- Since, we are considering scores out of 10, 3.5/5 will be treated as 7/10. …(3)

K. Artificial Neural Networks

Artificial Neural Networks are methods of computation resembling neural networks of biology which have the capacity to learn.

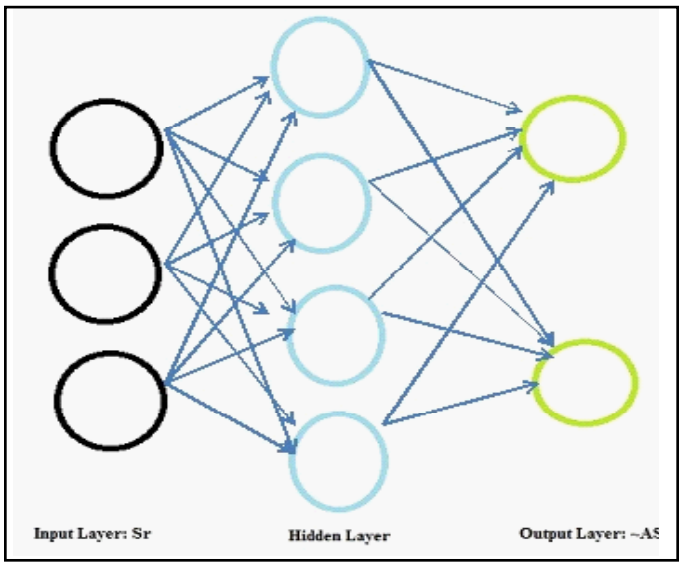

Fig.1: Artificial Neural Networks

represents number; U=User no; Ur=Actual Score/User Rating given by the reviewer; T=Title; Sg=Suggestions; UR=User Review; Sr=Suggested Rating/Score; %=percentage; A=Accuracy; Apartment=APARTMENT 1303 3D; Godfather=The Godfather; ANN=Artificial Neural Networks.

V. Dataset at a Glance

Our dataset consists of more than 1000 reviews collected from the website of ID. Let's look at such examples:-

Movie name: - "The Shawshank Redemption."
Title of the Review: - "Tied for the best movie I have ever seen."
Full Review: - "The only other movie I have ever seen that affects me as strongly is To Kill a Mockingbird. Both movies leave me feeling cleaner for having watched them. "

Movie name: - "Inception"
Title of the Review: - "Too much....WAY too much"
Full Review: - "What is going on with the ID user reviews lately? It's like the masses can no longer be trusted. In the last month, the users have decreed "Airbender" the worst abomination ever, when in fact it's just an average movie."
Note: In some cases where the reviews are too long, we have used a part of it for analysis.

VI. Workflow Diagram

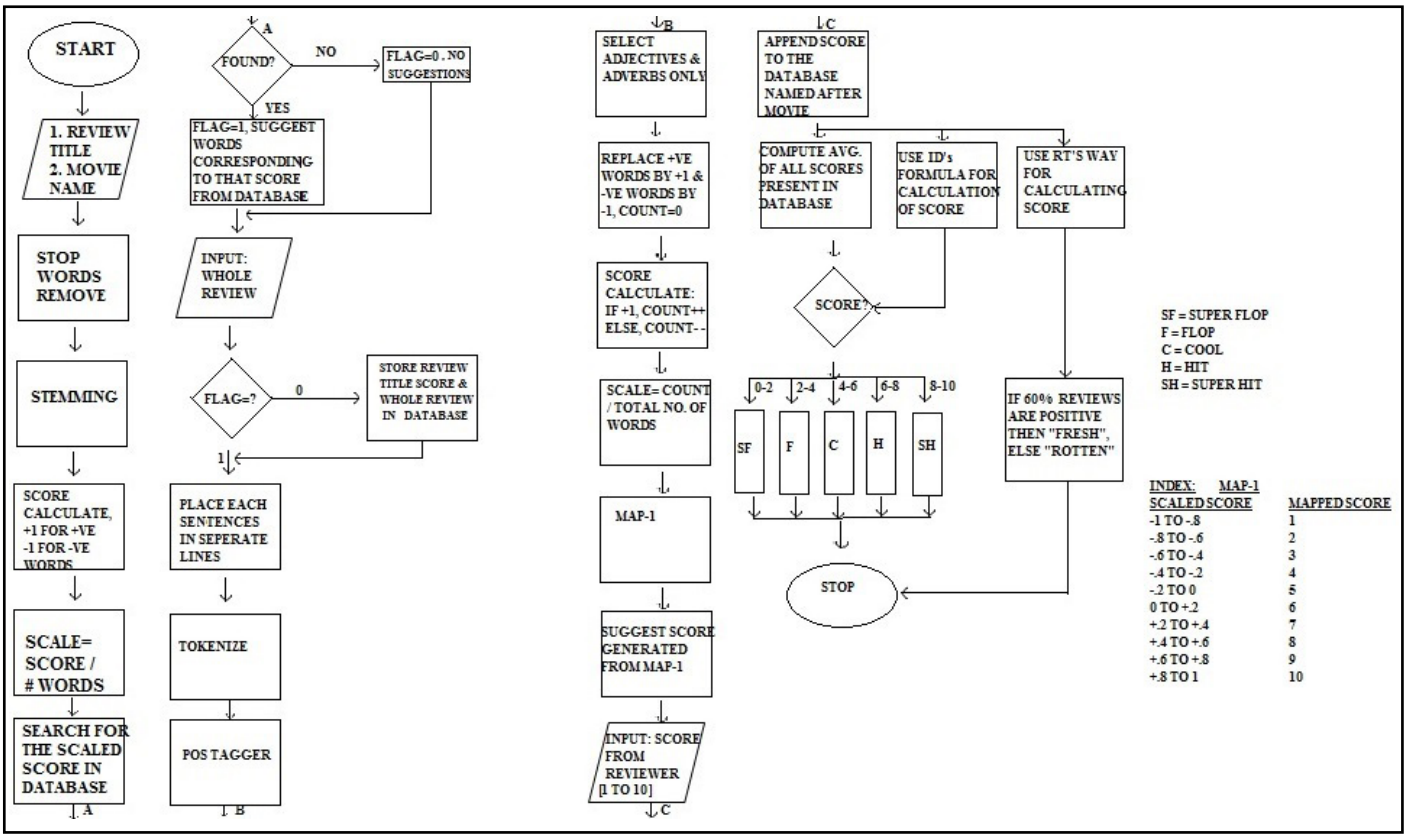

Fig. 2: Workflow Diagram of The Model

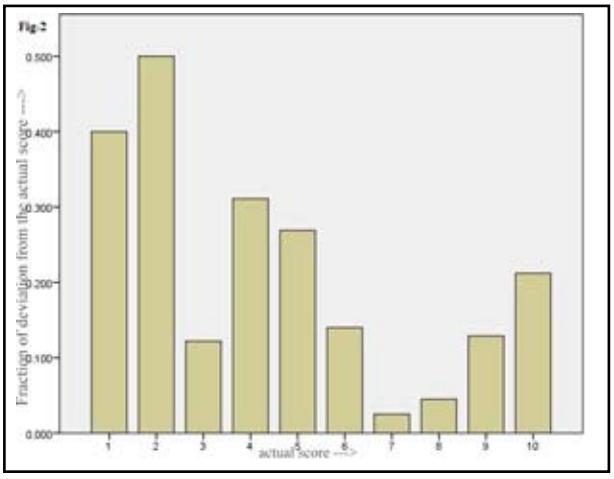

Fig. 3 : Plot between Ur VS FA

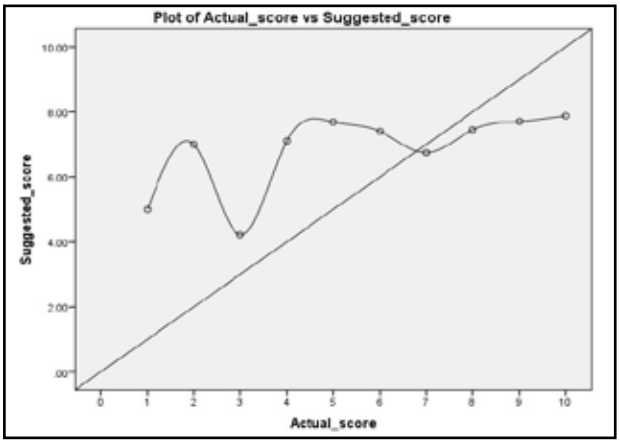

Fig. 4 : Plot between Ur VS Sr

VII. Experimental Evaluation

We have used our model to suggest words and compute the popularity of some films. Here are some of the instances: -

- Case-1:

Table: 1.Movie: Inception

U	T	Sg	UR	Sr	Ur
1	Too much.... WAY too much	average, ordinary, so-so, mediocre	What is going on with the ID user reviews lately? It's like the masses can no longer be trusted. In the last month, the users have decreed "Airbender" the worst abomination ever, when in fact it's just an average movie.	6	7

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

2	Insanely Brilliant! Nolan has outdone himself!!	first, disparate, only, brief, new, very, acting, so, well, easily, practically etc.	What is the most resilient parasite? An Idea! Yes, Nolan has created something with his unbelievably, incredibly and god-gifted mind which will blow the minds of the audience away.	10	10

• Case-2:

Table 2 : Movie: Apartment 1303 3D

U	T	Sg	UR	Sr	Ur
1	Hard to describe how bad this is!	only, negative, never, can't, dull, boring, probably, aren't, far, outweigh, rarely, past, traumatic,	It's one of those films where you realize you are in trouble about five minutes in. The acting is wooden and not believable. The writing and direction are not at a professional level, to put it nicely. The plot is ridiculous and there are no scares.	4	1
2	Waste of time!!	Sorry!! no suggestions for u!!	Where to start. This movie was crap. The acting is horrible. That girl cannot act. She sounds mannish and so mono toned throughout the movie.	1	1

• Case-3:

Table 3: Movie:the Godfather

U	T	Sg	UR	Sr	Ur
1	Excellent Movie	wonderful, very, best, little, probably, too, much, superb, outstanding, last, absolutely, naturalistic, best	The Godfather is without a doubt one of the best movies I've ever seen. I'm going to be one of those annoying people and say that the book is better than the movie, but the movie was still great. I've seen this movie multiple times and each time it gets	10	10
2	Great movie ever watched	vivid, real, delightfully, broadly, great, emotional, really, interested, such, strong, personal, fabulous	I love this movie and all of the GF movies. I see something new every time I have seen it (countless, truly). The story of tragedy and (little) comedy that exists in this film is easily understood by people all over the world.	10	10

Note: Sr is the suggestive rating, predicted by our model. Refer to part [6] for details of the calculations. Ur is the rating given by the user corresponding to that review in ID's website.

VIII. Comparative Analysis

In this section we discuss about the efficiency of our model. Here, we analyse on the basis of the above two user's rating. Let's have a look at these tables first: -

Table 4: Movie: Inception

Model	Score	Popularity
ID	7.00	Hit
RT	100%	Fresh
Our Model	8.5	Superhit

Table 5: Movie: Apartment 1303 3D

Model	Score	Popularity
ID	6.99	Hit
RT	0%	Rotten
Our Model	2.5	Flop

Table 6: Movie: The God Father

Model	Score	Popularity
ID	7.00	Hit
RT	100%	Fresh
Our Model	10	Superhit

Note: Here the measure of popularity of a movie [for ID and our model] has been computed using a mapping function. Refer to part

[6] for details. For RT it has been done by calculating percentage of users who have rated the movie by 6 or more. The calculation of score for ID has been done using eqn. (2), and that of our model is done using eqn. (1).
From Table 4, 5 and 6 we find that the ID rating is mostly same in all the cases. The RT score is 100% in case of Table 4 and 6. Thus, we can say for movies with lower # of ratings, the scoring process of RT is not very efficient. Our model is a better fit for these kinds of movies (like the regional ones). This is a disadvantage. In such situations, our model comes to a rescue.
Apart from this our model suggests words to the appraiser while reviewing a film. It proposes the most probable score to him. These innovative features are not found in traditional reviewing sites. If, these features are put to use, the conventional appraising method will be faster. This is because, a reviewer don't have to spend much time thinking for words he wants to use or the score he wants to assign.

Table 7: Actual Ratings

MOVIE	ACTUAL ID RATINGS	ACTUAL RT AUDIENCE RATINGS
INCEPTION	8.8	91%
APARTMENT 1303 3D	2.6	9%
THE GODFATHER	9.2	98%

Note: Actual ratings are noted from the respective websites.
Now, let's talk about the % efficiency of our model

Table 8: % Accuracy

MOVIE	USER #	Sr	Ur	%A
Inception	1	6	7	90%
Inception	2	10	10	100%
Apartment	1	4	1	70%
Apartment	2	1	1	100%
Godfather	1	10	10	100%
Godfather	2	10	10	100%

$$\%A = 100 - \left(\frac{|Sr-Ur|}{10}\right) * 100 \quad \ldots (4)$$

Average % accuracy for our model = {Σ (%A)}/n = 93.33%.
% accuracy refers to the percentage with which our model is capable of predicting score. It is the measure of efficiency with which it can anticipate score from the opinion. n = number of Movies for which we are computing the score. It's recommended to compute the % accuracy as many movies as possible. Here, we are showing only for 3 well-known films.
Let's look at the model from another perspective.
In Fig-3 we see the plot between the actual score (Ur) and the fraction of deviation from the actual score (FA). Fig-4 is the plot between Ur and Suggested Score(Sr). Here the x=y line represents the plot that should have occurred in ideal case. The curved line is the line obtained by interpolating the suggested scores..

$$FA = \frac{|Sr-Ur|}{10} \quad \ldots . (5)$$

Ur = score given by user [1-10]; we can observe from Fig. 3 that for score=7, our model can suggest most efficiently, for 2 it is least efficient and so on. Thus, there exists a scope to fine-tune it further.
Note: - For a particular actual score (Ur) we observe the score suggested by the model (Sr_i) multiple times. Finally, we took mean of all Sr_i to get Sr.

IX. Fine-Tuning The Model

It is possible to fine-tune our innovative model further. We use Artificial Neural Networks (ANN) to maximize its efficiency. Let's look how it is done:-
Fig-1 illustrates the working of ANN. We give the score generated by our model (Sr), deviation FA and other parameters* as input. The hidden layers perform the necessary computations and train it. The output score is noted. Necessary changes are done to make it efficient. Finally when we use it for our test set, its outcome nearly matches with that of Actual Score (Ur). Thus, we can minimize the deviation (FA). Each stage of the hidden layers have certain biases Θ which transforms the input value into the desired one.
*Note: These parameters are dependent on the type of the model we want to use to anticipate the score. For ID and RT the parameters will be different.

X. Conclusion And Future Work

In this paper we discuss about our unique model to propose words to a reviewer while appraising a film. We use the online database of movies, ID to build our training and test set. Firstly, we request the appraiser to enter the name of the movie and the title of his opinion. We analyze this title and predict words he is most likely to use while reviewing. Then, we give him a turn to opine. We examine this opinion and compute score from it. Furthermore, we give him a chance to rate the film. We repeat this process for every user. For each movie we store every appraiser's rating. We evaluate the mean from this and declare whether the movie is 'super-flop', 'flop', 'cool', 'hit' or 'super-hit'. Finally, we make a comparative study of ratings given by ID and RT.
This model is beneficial as it makes the process of appraising faster and simpler. Users do not have to spend time wondering for words while reviewing a film. They don't need to consider about the score they want to assign. They will receive suggestions at each and every step. Moreover, this model has the ability to learn which enhances its efficiency with usage.
This ingenious model is quite handy, fast and fit for use. It saves time. There are few scopes of improvement. We have trained it using about 1000 reviews. It's advisable to train it with more reviews to enhance its accuracy. Here, we check the polarity of a word by checking its presence in positive or negative corpora. Based on the scores suggested by our model, a recommendation system can be built which will recommend movies to the users. Instead of suggesting too many words, it is better to suggest those with higher frequency of occurrences.

Reference

[1] *Sima, I.M.; Vunvulea, M., "A rule-based, domain independent approach for opinion and holder identification," Intelligent Computer Communication and Processing (ICCP), 2013 IEEE International Conference on , vol., no., pp.55,62, 5-7 Sept.*
[2] *Rustamov, S.; Mustafayev, E.; Clements, M.A., "An application of hidden Markov models in subjectivity analysis," Application of Information and Communication Technologies (AICT), 2013 7th International Conference*

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

on , vol., no., pp.1,4, 23-25 Oct. 2013

[3] *Jusoh, S.; Alfawareh, H.M., "Applying fuzzy sets for opinion mining," Computer Applications Technology (ICCAT), 2013 International Conference on , vol., no., pp.1,5, 20-22 Jan. 2013*

[4] *Weiping Wang; Yuanzhuang Zhou, "E-business Websites Evaluation Based on Opinion Mining," Electronic Commerce and Business Intelligence, 2009. ECBI 2009. International Conference on , vol., no., pp.87,90, 6-7 June 2009*

[5] *Kamal, A.; Abulaish, M., "Statistical Features Identification for Sentiment Analysis Using Machine Learning Techniques," Computational and Business Intelligence (ISCBI), 2013 International Symposium on , vol., no., pp.178,181, 24-26 Aug. 2013*

[6] *Rustamov, S.; Mustafayev, E.; Clements, M.A., "An application of hidden Markov models in subjectivity analysis," Application of Information and Communication Technologies (AICT), 2013 7th International Conference on , vol., no., pp.1,4, 23-25 Oct. 2013*

[7] *Maas, Andrew L. et al "Learning Word Vectors for Sentiment Analysis" in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 2011.*

[8] *Jacob Perkins, "Python Text Processing with NLTK 2.0 Cookbook", PACKT publishing*

[9] *http://www.enchantedlearning.com/wordlist/positivewords.shtml, http://www.enchantedlearning.com/wordlist/negativewords.shtml; List of positive and negative words.*

[10] *http://computer-ease.com/emotposi.htm, http://computer-ease.com/emotneg.htm List of positive and negative emoticons.*

[11] *http://www.nltk.org/*

[12] *Wong, Felix Ming Fai, Soumya Sen, and Mung Chiang. "Why watching movie tweets won't tell the whole story?." Proceedings of the 2012 ACM workshop on Workshop on online social networks. ACM, 2012*

[13] *Sing, J. K., Souvik Sarkar, and Tapas Kr Mitra. "Development of a novel algorithm for sentiment analysis based on adverb-adjective-noun combinations." Emerging Trends and Applications in Computer Science (NCETACS), 2012 3rd National Conference on. IEEE, 2012*

[14] *M.F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130−137*

[15] *Bing Liu. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, May 2012M.*

Author Profiles

Mr. Sohom Ghosh is a student at Heritage Institute of Technology, Kolkata. He is pursuing B.Tech in Computer Science and Engineering. His research interests include Data Mining, Social Network Analysis and Machine Learning.

Santanu Modak is Junior Research Fellow in the Department of Computer Science, University of Burdwan. He received his B.Sc(Hons) and M.Sc in Computer Science degrees in 2009 and 2011 respectively. He qualified UGC NET in Computer Science and Applications. He published four international journals. He is a life member of Indian Science Congress Association (ISCA), member of International Association of Computer Science and Information Technology (IACSIT) and member of International Association of Engineers.

Abhoy Chand Mondal is currently Associate Professor of Department of Computer Science, Burdwan University, W.B., India. He received his B.Sc.(Mathematics Hons.) from The University of Burdwan in 1987, M.Sc. (Math) and M.C.A. from Jadavpur University, in 1989, 1992 respectively. He received his Ph.D. from Burdwan University in 2004. He has 1 year industry experience and 18 years of teaching and research experience. No. of papers more than 50 and no of journal published is 25.