Simultaneous Prediction of Stock Market Investments by Analyzing Sentiments: A Supervised Joint Aspect Model

[1]Chethan Chandra, [2]S Basava Raddy, [3]Tejaswini P R, [4]Yashaswini P R

[1,2]Assistant Professor, [3,4]M.tech, 4th sem, CSE,

KIT,Tiptur-572201

Abstract-Sentimental Analysis is one of the most popular technique which is widely been used in every industry. Extraction of sentiments from user's comments is used in detecting the user view for a particular company. Sentimental Analysis can help in predicting the mood of people which affects the stock prices and thus can help in prediction of actual prices. In this paper sentimental analysis is performed using supervised joint aspect and sentiment model (SJASM) on the data extracted from Twitter and Stock Twits. The data is analysed to compute the mood of user's comment. These comments are categorized into four categories which are happy, up, down and rejected. The novel probabilistic supervised joint aspect and sentiment model (SJASM) deal with the problems in one go under a unified framework. SJASM represents each review document in the form of opinion pairs, and can simultaneously model aspect terms and corresponding opinion words of the review for hidden aspect and sentiment detection. There by sentiments are analysed to predict which company is good stock investments.

Keywords: Sentiment analysis, Stock, SJASM, Twits, aspect-based sentiment analysis.

I. INTRODUCTION

Sentimental Analysis also known as Opinion Mining is an area that uses Natural Language Processing and Text Analysis that helps in building a system that identifies and extract information in source material. An initial task in sentimental analysis is to determine the polarity of a specified text at the document level, sentence level or aspect level. In core, it is a process that helps in determining the emotional level behind a sequence of words, used to gain an insight of speaker's attitude, opinions and emotions expressed in a sentence.

Online user-generated reviews are of great practical use, because: 1) they have become an inevitable part of decision making process. 2) they collectively form a low-cost and efficient feedback channel, which helps in making decisions regarding purchasing and selling of stocks in market. As a matter of fact, online reviews are constantly growing in quantity, while varying largely in content quality. To support users in digesting the huge amount of raw review data, many sentiment analysis techniques have been developed for past years.

Sentimental Analysis is very useful in social media monitoring as it provides an insight of public opinions for certain topics. The uses of sentimental analysis are very extensive and powerful. Sentimental Analysis provides the ability to extract insight from social data which is broadly used by various organizations across the world. Prediction of the market and stock prices for a company has always been a wide area for the researchers to work upon. A company can be successful in long run only if its consumers are happy with its performance and are giving positive feedback for its products. Expedia Canada used this technique to quickly understand consumer attitude which increased in a negative feedback towards one of their television advertisements. Sentimental Analysis is a widely used field giving many benefits to every industry. Thus if sentiments are correctly categorized and their polarity are correctly determined they can be helpful in enhancing a company's performance and making its investors happy.

II. SYSTEM MODEL

2.1 PROBLEM STATEMENT

When customers invest into stock market he can lose a quite amount of money if he didn't understands how share prices are fluctuating. In the simplest sense, investors buy shares at a certain price and can then sell the shares to realize capital gains. However, the investor will not realize a gain if the share price drops dramatically; in fact, the investor will lose money. Hence an investor can lose a large amount of money if he has no prior knowledge of how stock prices are changing. So by getting knowledge by reviews customers can save themselves from losses and can even earn profit.

Since we analyze user-generated review data, we first provide the definitions of the terminologies commonly used in the sentiment analysis of user reviews.

Aspect Term: An aspect term t, also known as feature or explicit feature, indicates a specific attribute or component word of an opinionated entity, which typically appears as noun or noun phrase in review text. For instance, the noun "trading" is an aspect term in review, "TD Ameritrade is best for trading"

Opinion Word: An opinion word o, also called sentiment word, refers to the word used to express subjectivity or sentiments, and typically appears as an adjective in review documents. For example, the word “best” is recognized as an opinion word from the aforementioned example review.

Opinion Pair: An opinion pair op = ⟨t; o⟩ is simply defined as a pair of aspect term t and corresponding opinion word o extracted from a given review document. For instance, one opinion pair op = ⟨trading; best⟩ can be recognized from the example review above. The extracted opinion pairs would constitute the input to our sentiment analysis system.

Aspect: An aspect a, or semantic aspect, refers to a unique facet that corresponds to a rateable attribute or component of an opinionated entity. In our setting, it is formulated as latent variable, and is typically represented as a hidden cluster of semantically related aspect terms or opinion words. For instance, one can detect a rateable semantic aspect platform from the following reviews, “TD Ameritrade is affordable platform” and “It is really a discount broker.”

Sentiment: A sentiment s, or opinion, refers to the semantic orientation and degree or strength of satisfaction on a reviewed entity or its aspect in a review text. Positive semantic orientation indicates praise i.e., Happy or Up , while negative semantic orientation indicates criticism i.e, down or rejected. In our setting, sentiment is formulated as a latent variable, and refers to a hidden semantic cluster of opinion words which share the same sentimental polarity.

Overall Rating: The overall rating r indicates the degree of sentiment demonstrated in a whole review document. Then, given an entity from a company, there is a collection of M review documents on the entity, D = $\{d_1,d_2,\ldots\ldots,d_M\}$.Each review d_m can be reduced to a list of N opinion pairs: d= $\{<t_1,o_1>,<t_2,o_2>,<t_3,o_3>\ldots\ldots.<t_N,o_N>\}$ ⟩, where each opinion pair consists of an aspect term t word o and corresponding opinion word o_n in the review. We aim to deal with three sentiment analysis tasks as follows.

Semantic aspect detection: This task aims at detecting hidden semantic aspects of an opinionated entity from the given review documents, where each aspect would be represented in the form of a hidden semantic cluster.

Aspect-level sentiment identification: For this task, the aim is to identify fine-grained semantic sentiment orientations, e.g., positive or negative, expressed towards each detected semantic aspect.

Overall rating/sentiment prediction: Given an unlabelled review, we will form the prediction for the overall sentimental rating by employing a carefully designed regression procedure over the inferred hidden aspects and aspect-level sentiments via the fitted model.

The novel formulation behind the proposed SJASM model actually agrees well with intuitions. Generally, different Stock investments have diverse lists of aspects, e.g., attributes or components. The utility quality of individual product aspects could be different, and may result in different evaluations and opinions on the aspects. Overall experiences and sentiments on the products would be formed or regressed on the product aspects and their associated evaluations expressed in the reviews. The regression coefficients reflect the relative contributions of the fine-grained aspect-specific sentiments. Furthermore, given user-generated review and rating pair data, the labeled overall ratings of review documents can be leveraged as supervision knowledge. They thus provide useful guidance and constraint on the procedure of inferring the meaningful and predictive hidden aspects and sentiments. In addition to overall rating data, SJASM also leverages a pre-compiled sentiment lexicon as weak supervision information, which not only benefits semantic sentiment analysis, but also provides explicit correspondence between latent sentiment variables and real-world sentiment orientations (e.g., happy, up, down and rejected).

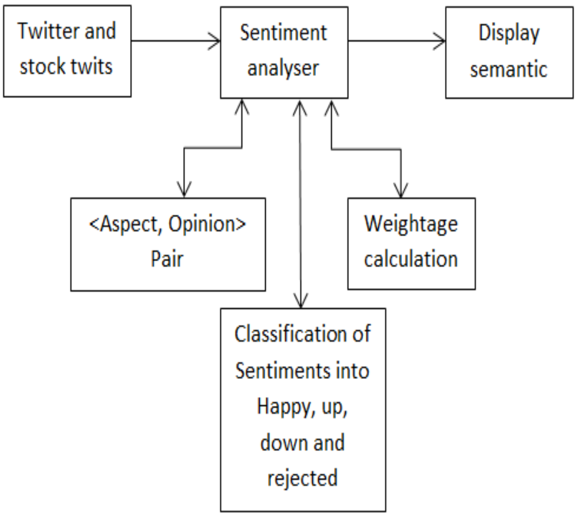


Fig. 2.1 System Architecture

III. PREVIOUS WORK

During past years much work has been contributed by the researchers on Sentiment analysis. Pang et al. [1] classified documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, they find that standard machine learning techniques definitively outperform human-produced baseline. they proposed supervised models on standard n-gram text features to classify review documents into positive or negative sentiments and also considered Naive Bayes, maximum entropy classification, and support vector machines for traditional topic-based categorization.

V. Ng, S. Dasgupta, and S. M. N. Arifin [2] worked on to examine two problems in document-level sentiment analysis: (1) determining whether a given document is a review or not, and (2) classifying the polarity of a review as positive or negative. Using only unigrams as features they demonstrated that review identification can be performed with high accuracy.

J. Zhao, K. Liu, and G. Wang [3] discussed a novel method based on CRF's on two special characteristics of "contextual dependency" and "label redundancy" in sentence sentiment classification. They introduced redundant labels into the original sentimental label set and organized all of them into a hierarchy; this method can add redundant features into training for capturing the label redundancy.

P. Melville, W. Gryc, and R. D. Lawrence [4] analysed overall sentiments of blog and review documents, by incorporating background/prior lexical knowledge based on a pre-compiled sentiment lexicon into a supervised pooling multinomial text classification model.

Phayung Meesad, Jiajia Li.[5] developed a text-sentiment based stock trend prediction model with a hybrid feature selection method. They used SentiWordNet to give an additional weight to the selected features.

Z. Hai, K. Chang, G. Cong, and C. C. Yang [6] employed a corpus statistics association measure to quantify the pairwise word dependencies and proposed a generalized association-based unified framework to identify features, including explicit and implicit features, and opinion words from reviews. Very first they extracted explicit features and opinion words via an association-based bootstrapping method (ABOOT). Two instances of this ABOOT method are evaluated based on two particular association models, likelihood ratio tests (LRTs) and latent semantic analysis (LSA).

Sprenger and Webye [7] adopted an automatically dictionary construction approach and sentiment analysis of stock market news using the dictionary. They compare polarities determined by a financial expert with polarities determined with proposed method. There by they revealed that proposed method can make an appropriate dictionary.

Hui Song, Yingxiang Fan, Xiaoqiang Liu and Dao Tao[8] proposed an approach based on patterns to extract features. Trough setting length, upper and lower limit probability and frequency thresholds, they extracted patterns of positive tags and features from the training corpus. To enhance adaptability of the pattern set, they merge some fundamental patterns into a new fuzzy pattern. Then they applied a pattern matching algorithm to extract the titles and opinion words from the reviews.Thomas P Minka [9] explained that the Dirichlet distribution and its compound variant, the Dirichlet-multinomial, are two of the most basic models for proportional data, such as the mix of vocabulary words in a text document. Yet the maximum-likelihood estimate of these distributions is not available in closed-form. This paper describes simple and efficient iterative schemes for obtaining parameter estimates in these models. In each case, a fixed-point iteration and a Newton-Raphson or generalized Newton-Raphson iteration is provided.

Neethu M.S and Rajasree R. [10] analyzed the twitter posts about electronic products like mobiles, laptops etc using Machine Learning approach. By doing sentiment analysis in a specific domain, made possible to identify the effect of domain information in sentiment classification. They presented a new feature vector for classifying the tweets as positive, negative and extract peoples' opinion about products.

Sunil Kumar Khatri, Himanshu Singhal and Prashant Johri [11] formulated sentiment analysis on data from social media which is classified using classification algorithm of machine learning. The classified data was analysed to calculate the net mood of the comments. These comments were classified into four classes' namely happy, hope, sad, disappointing. The net relative mood of all the classes per day is used as input for artificial neural network (ANN) to be trained for data of n days and their respective change in index value on each day. This trained network was finally used to predict the vector of Bombay Stock Exchange index value for $(n+1)^{th}$ days.

Sunil Kumar Khatri and Ayush Srivastava [12] analysed sentiments by extracting data from Twitter and Stock Twits. They analysed the data to compute the mood of user’s comment. These comments are categorized into four categories which are happy, up, down and rejected. The polarity index along with market data was supplied to an artificial neural network to predict the results.

Zhen Hai, Gao Cong, Kuiyu Chang, Peng Cheng, and Chunyan Miao [13] focused on modeling user-generated review and overall rating pairs, and aimed to identify semantic aspects and aspect-level sentiments from review data as well as to predict overall sentiments of reviews using novel probabilistic supervised joint aspect and sentiment model (SJASM) to deal with the problems simultaneously.

IV. PROPOSED METHODOLOGY

4.1 OVERVIEW

We model online user-generated review and overall rating pairs, and aim to identify semantic aspects and aspect-level sentiments from review texts as well as to predict overall sentiments of reviews.

User-generated reviews are different from ordinary text documents. For example, when people read a product review, they often care about which specific aspects of the product are commented on, and what sentiment orientations (Up, Down, Happy, Rejected) have been expressed on the aspects. Instead of employing bag-of-words representation, which is typically adopted for processing usual text documents, we represent each review in an intuitive form of opinion pairs, where each opinion pair consists of an aspect term and related opinion word in the review. Probabilistic topic models, notably latent Dirichlet allocation (LDA), have been widely used for analyzing semantic topical structure of text data. Based on the basic LDA, we introduce an additional aspect-level sentiment identification layer, and construct a probabilistic joint aspect and sentiment framework to model the textual bag-of-opinion-pair data. Online user-generated reviews often come with overall ratings (sentiment labels), which provides us with great flexibility to develop supervised unification topic model.

4.2 Supervised Joint Aspect and Sentiment Model

We make the following assumptions about our proposed SJASM model:

- The generation for aspect-specific sentiments depends on the aspects. This means that we first generate latent aspects, on which we subsequently generate corresponding sentiment orientations.
- The generation for aspect terms depends on the aspects, while the generation for opinion words relies on the sentiment orientations and semantic aspects. The formulation is intuitive, for example, to generate an opinion word “best”, we need to know its sentiment orientation positive and related semantic aspect platform.
- The generation for overall ratings of reviews depends on the semantic aspect-level sentiments in the reviews.

Based on the model assumptions, to generate a review document and its overall rating, we first draw hidden semantic aspects conditioned on document-specific aspect distribution; We then draw the sentiment orientations on the aspects conditioned on the per document aspect-specific sentiment distribution; Next, we draw each opinion pair, which contains an aspect term and corresponding opinion word, conditioned on aspect and sentiment specific word distributions; We lastly draw the overall rating response based on the generated aspect and sentiment assignments in the review document.

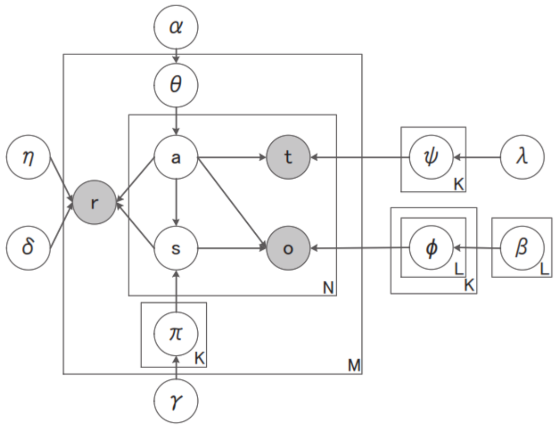


Fig. 4.1 Graphical representation of SJASM

The boxes refer to plates that indicate replicates. The outer plate refers to review documents, while the inner plate refers to the repeated selection of latent aspects and sentiment orientations as well as aspect terms and opinion words within each review document.

The graphical model representation of SJASM is shown in Figure 4.1, and the notations used in this model are listed in figure4.2. The generative process of the graphical model is as follows:

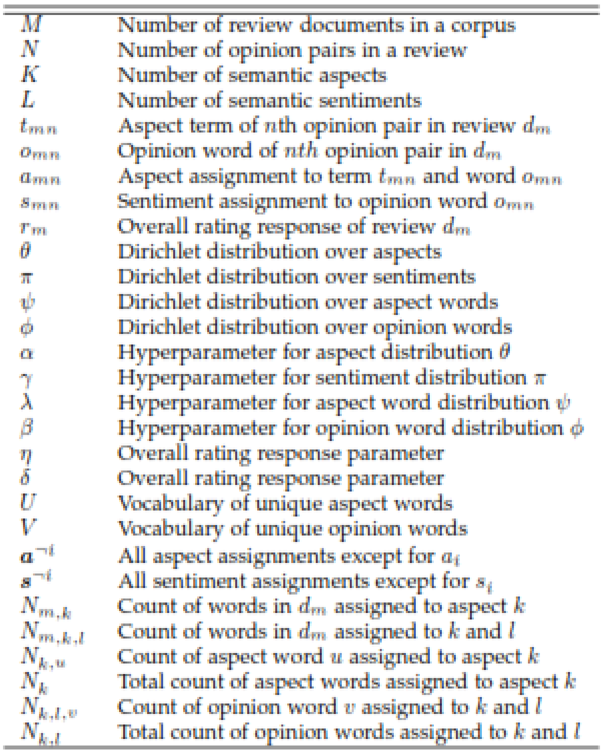
1. For each aspect k ∈ {1, . . . , K}
(a) Draw aspect word distribution $\psi k \sim Dir(\lambda)$.
(b) For each sentiment orientation l ∈ {1, . . . , L}.
i. Draw opinion word distribution $\varphi_{kl} \sim Dir(\beta_l)$.

2. For each review d_m and its helpfulness response r_m
(a) Draw aspect distribution $\theta_m \sim Dir(\alpha)$.
(b) For each aspect k under review r_m
i. Draw sentiment distribution $\pi_{mk} \sim Dir(\gamma)$.
(c) For each opinion pair $<t_{mn}, o_{mn}>$, n ∈ {1, . . . , N}
i. Draw aspect $a_{mn} \sim Mult(\theta_m)$.
ii. Draw sentiment $s_{mn} \sim Mult(\pi ma_{mn})$.
iii. Draw aspect term $t_{mn} \sim Mult(\psi a_{mn})$.
iv. Draw opinion word $o_{mn} \sim Mult(\varphi a_{mn})$.
(d) Draw overall rating response $r_m \sim N(n^T \bar{z}_m , \delta)$

$\bar{z}_m$ refers to the empirical frequencies of hidden variables or latent aspects and sentiments in the review document d_m , and is defined as

$$\bar{z}_m = \frac{1}{C}\sum_{n=1}^{N}(a_{mn} \times (\omega^T \times s_{mn})),$$

…………….. (1)



M	Number of review documents in a corpus
N	Number of opinion pairs in a review
K	Number of semantic aspects
L	Number of semantic sentiments
t_{mn}	Aspect term of nth opinion pair in review d_m
o_{mn}	Opinion word of nth opinion pair in d_m
a_{mn}	Aspect assignment to term t_{mn} and word o_{mn}
s_{mn}	Sentiment assignment to opinion word o_{mn}
r_m	Overall rating response of review d_m
θ	Dirichlet distribution over aspects
π	Dirichlet distribution over sentiments
ψ	Dirichlet distribution over aspect words
ϕ	Dirichlet distribution over opinion words
α	Hyperparameter for aspect distribution θ
γ	Hyperparameter for sentiment distribution π
λ	Hyperparameter for aspect word distribution ψ
β	Hyperparameter for opinion word distribution ϕ
η	Overall rating response parameter
δ	Overall rating response parameter
U	Vocabulary of unique aspect words
V	Vocabulary of unique opinion words
$\boldsymbol{a}^{\neg i}$	All aspect assignments except for a_i
$\boldsymbol{s}^{\neg i}$	All sentiment assignments except for s_i
$N_{m,k}$	Count of words in d_m assigned to aspect k
$N_{m,k,l}$	Count of words in d_m assigned to k and l
$N_{k,u}$	Count of aspect word u assigned to aspect k
N_k	Total count of aspect words assigned to aspect k
$N_{k,l,v}$	Count of opinion word v assigned to k and l
$N_{k,l}$	Total count of opinion words assigned to k and l

Fig. 4.2: Notations used in SJASM

The novel formulation behind the proposed SJASM model actually agrees well with intuitions. Generally, different products have diverse lists of aspects, e.g., attributes or components. The utility quality of individual product aspects could be different, and may result in different evaluations and opinions on the aspects. Overall experiences and sentiments on the products would be formed or regressed on the product aspects and their associated evaluations expressed in the reviews. The regression coefficients reflect the relative contributions of the fine-grained aspect-specific sentiments. Furthermore, given user-generated review and rating pair data, the labeled overall ratings of review documents can be leveraged as supervision knowledge. They thus provide useful guidance and constraint on the procedure of inferring the meaningful and predictive hidden aspects and sentiments.

In addition to overall rating data, SJASM also leverages a pre-compiled sentiment lexicon as weak supervision information, which not only benefits semantic sentiment analysis, but also provides explicit correspondence between latent sentiment variables and real-world sentiment orientation(e.g., positive or negative).

User-generated review data are different from usual textual articles. When people read reviews, they typically concern themselves with what aspects of an opinionated entity are mentioned in the reviews, and which sentiment orientations are expressed towards the aspects. Thus, instead of using traditional bag-of-words representation, we reduce each text review as a bag of opinion pairs, where each opinion pair contains an aspect term and related opinion word appearing in the review. Specifically, we parsed all the text reviews in each data set using the well-known Stanford Parser, We then straightforwardly use the syntactic dependency patterns to recognize the opinion pairs from the review texts. As a separate preprocessing step, several other methods, which were specially developed for extracting aspect terms and corresponding opinion words from reviews may also work in generating the bagof-opinion-pairs representation. It's true that better opinion pair extraction results would be beneficial for the proposed model SJASM to lead to improved performance of the sentiment analysis tasks. The proposed SJASM model belongs to the family of generative probabilistic topic models to sentiment analysis. SJASM is able to model the hidden thematic structure of text review data, and thus, similar to other unsupervised or weakly supervised joint topic-sentiment (sentiment-topic) models, it can rely on per document-specific sentiment distribution to approximate the overall sentiments of text reviews.

V. CONCLUSION

Sentimental Analysis provides the ability to analyze the opinions of people for a particular product or for a company. Prediction of stock market is really a hard nut to crack and requires lot of efforts. The market data if analyzed in a proper way can be very effectual in predicting a company's future. Stock investment prediction are made to analyse in one go. Online stock reviews generated by user are used in prediction which helps customers in investing in different companies. These predictions are done by using a novel supervised joint aspect and sentiment model (SJASM) to deal with the problems in one go under a unified framework. SJASM treats review documents in the form of opinion pairs, and can simultaneously model aspect terms and their corresponding opinion words of the reviews for semantic aspect and sentiment detection. Moreover, SJASM also leverages overall ratings of reviews as supervision and constraint data, and can jointly infer hidden aspects and sentiments that are not only meaningful but also predictive of overall sentiments of the review documents.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, ser. EMNLP'02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86.

[2] V. Ng, S. Dasgupta, and S. M. N. Arifin, "Examining the role of linguistic knowledge sources in the automatic

identification and classification of reviews," in Proceedings of the COLING/ACL on Main Conference Poster Sessions, ser. COLING-ACL '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 611–618.

[3] J. Zhao, K. Liu, and G. Wang, "Adding redundant features for crfs-based sentence sentiment classification," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 117–126.

[4] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD'09. New York, NY, USA: ACM, 2009, pp. 1275–1284.

[5] Phayung Meesad, Jiajia Li. Stock trend prediction relying on text mining and sentiment analysis with tweets, Fourth World congress on Information and Technology,2014,pp.257-262.Name of Author(s), "Title of the research", Citation Details, year.

[6] Z. Hai, K. Chang, G. Cong, and C. C. Yang, "An association-based unified framework for mining features and opinion words," ACM Trans. Intell. Syst. Technol., vol. 6, no. 2, pp. 26:1–26:21, Mar. 2015.

[7] Sprenger and Webye, "Sentiment Analysis of Stock Market News with Semi-supervised Learning", International Conf. on Computer and Information Science, 2012, pp. 325-328.

[8] Hui Song,Yingxiang Fan,Xiaoqiang Liu and Dao Tao, "Extracting product features from online reviews for sentimental analysis", Computer Science and Converge Information Technology,2014, pp. 741-750.

[9] T. Minka, "Estimating a dirichlet distribution," in Technical report, MIT, 2000.

[10] Neethu M.S and Rajasree R., "Sentiment Analysis in Twitter Using Machine Learning Technique", 4 ICCC NT, 2013, pp. 1-5.

[11] Sunil Kumar Khatri, Himanshu Singhal and Prashant Johri, "Sentimental analysis to Predict Bombay Stock Exchange Using Artificial Neural Network", Proc. Of ICRITO,2014, pp. 380-384.

[12] Sunil Kumar Khatri , Ayush Srivastava, "Using Sentimental Analysis in Prediction of Stock Market Investment", 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Sep. 7-9, 2016, AIIT, Amity University Uttar Pradesh, Noida, India.

[13] Zhen Hai, Gao Cong, Kuiyu Chang, Peng Cheng, and Chunyan Miao "Analyzing Sentiments in One Go: A Supervised Joint Topic Modeling Approach", IEEE Transactions on Knowledge and Data Engineering (Volume: 29, Issue: 6, June 1 2017).