ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

Enhanced Interpretation of the Public Sentiment Variations on Social Networking using Improved LDA-based Approaches

Osamah A.M Ghaleb[1] and Anna Saro Vijendran[2]
[1]Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science
Coimbatore 641006, India
[2]Dean, School of Computing, Sri Ramakrishna College of Arts and Science
Coimbatore 641006, India

Abstract: Sentiment variation analysis has become an interesting area of research especially for the business marketing of the organizations. The decision making of major concerns and service providers focus on these sentiment analysis models for better user understanding. Latent Dirichlet Allocation (LDA) is a popular analysis model from which Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA) models have been developed for interpreting sentiment variation in the public statements in social networks. However, these models has inference problem which has been resolved previously using Gibbs Sampling. Still the issue of slow convergence in Gibbs sampling reduces the overall performance efficiency. Hence in this paper, a hybridized Collapsed Gibbs Sampling (CGS) and Collapsed Variational Bayes (CVB) (HCGS-CVB) method is proposed to reduce the inference problem within proved convergence rate. The HCGS-CVB method is used in both FB-LDA and RCB-LDA models to form two new models referred as FB-HCGS-CVB-LDA and RCB-HCGS-CVB-LDA respectively. The first model distills the foreground topics by filtering out the background topics while the most representative tweets are selected by the second model using ranking process based on tweet importance. For tracking public sentiment variation, sentiment analysis in different time interval is analyzed using improved TwitterSentiment and improved SentiStrength. The experiments are conducted on Twitter dataset collected using Twitter API shows that the proposed analysis model outperforms the existing models with higher accuracy and precision.

Keywords: Sentiment analysis, Sentiment variation tracking, Collapsed Gibbs Sampling, Collapsed Variational Bayes, Latent Dirichlet Allocation

I. INTRODUCTION

Textual information [1] is classified as facts and opinions. Facts are defined as objective expression about events, entities and their properties. Opinions are generally in the form of subjective expressions which defines people's feelings, sentiments or appraisals towards events, entities and their properties. Sentiment analysis [2] is a computational study of sentiments, emotions and opinions that are expressed in the text. The people feelings are expressed in positive, negative or neutral comments which are obtained by Natural Language Processing [3] and Information Extraction task [4]. There are different levels of sentiment analysis has been developed such as document level [5], sentence level [6], entity level and aspect level analysis [7]. The sentiment analysis provides the effective and economical way for exposing the public opinion that is utilized for decision making in different domains.

There have been different research studies and industrial applications in the area of public sentiment tracking [8] and modelling. But there is no research is performed to determine the causes of public sentiment variation effectively, which can be used to mine useful insights behind the sentiment variation and it can provide important decision making information. It is more important to find the possible reason behind the sentiment variation. Generally it is more difficult to find the exact causes of the sentiment variation. The emerging topics discussed in the variation period have been observed as more important reasons behind the sentiment variations. But mining emerging topics in twitter dataset is a challenging one because of it may contains noisy data and contain background topics that is discussed for a

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

long time and it cannot be a reason for the public sentiment variation [9].

As FB-LDA and RCB-LDA models which introduced by [9] has inference problem and slow convergence rate, this paper utilizes hybridized Collapsed Gibbs Sampling (CGS) and Collapsed Variational Bayes (CVB) (HCGS-CVB-LDA). Initially the improved TwitterSentiment and improved SentiStrength [10] are used to label the sentiments. The inference problem and slow convergence problem in LDA model is resolved by using a proposed effective hybridized CGS and CVB method. Based on the HCGS-CVB-LDA the foreground topics are distilled and filtered the background topics by using FB-HCGS-CVB-LDA. The most representative tweets for the filtered foreground topics are selected and ranked by using RCB-HCGS-CVB-LDA. Thus the proposed models have improved the performance of tracking and interpreting the public sentiment variations.

II. Related Work

Li et al [11] proposed an algorithm for sentiment analysis with the combination of Latent Dirichlet Allocation (LDA) model and a Gibbs sampling implementation. The combination of both LDA and Gibbs sampling was used to categorize the emotion tendency of people automatically with the help of computer. Moreover, a merge algorithm was proposed for computation of objects more accurately. However, in this method the context of the non-dependency of the emotion words affects the recall rate. Liang et al [12] introduced sentiment classification method called as Auxiliary Sentiment Latent Dirichlet Allocation (AS-LDA) method for sentiment analysis. The sentiment classification method considered the words in subjective documents consists of two parts are auxiliary words and sentiment analysis words. The AS-LDA method treat the auxiliary word vary from the sentiment element words. Gibbs sampling is used to perform the model inference.

Shams et al [13] proposed a method based on combining prior domain knowledge into the LDA topic model for aspect extraction. This method integrates the LDA model with co-occurrence of words in the document. An iterative algorithm is used where in each cycle based on co-occurrence prior knowledge from similar aspects are extracted. The knowledge from the iterative algorithm was added with the LDA model [14] in knowledge pair sets. Thus this process improves the quality of aspects. But the running time of this method is increased linearly which is the major drawback of this method.

Hai et al [15] proposed a novel probabilistic supervised joint aspect and sentiment model (SJASM) to identify semantic aspects and aspect level sentiments from review data and also for overall sentiment prediction. Furthermore, for the parameter estimation of SJASM an inference method called as collapsed Gibbs Sampling was developed. The major drawback of SJASM method is it cannot automatically estimate the number of latent topics from review data. Liu et al [16] extended LDA model to analysis behavior of online social network user. Then each topic was categorized based on both the multinomial distribution over words and Gaussian distribution over personality traits. Furthermore, a Gibbs Expectation Maximization algorithm was developed based on the Expectation maximization and Gibbs sampling that solves the iterative problem in LDA model. But this model doesn't consider both the topic and sentiments.

Shams et al [17] proposed a novel unsupervised LDA based sentiment analysis named as LDASA for sentiment analysis. Initially the Persian Clues was generated by an automatic translation approach which translates the English clues to Persian clues. The erroneous clues in the results of automatic translation approach were corrected by an iterative approach. Then these clues were used to obtain the topic based polar sets and the documents are categorized based on its polarity through Support Vector Machine (SVM) classification algorithm [18]. The SVM classification has disadvantages in size and speed of processing.

Onan et al [19] examined the performance of LDA model for text sentiment classification. LDA can be used as a viable method for representation of text collections in a compact yet effective way. The predictive performance of classification algorithms were improved by ensemble learning. Zhang et al [20] proposed a LDA based model named as CRATS which jointly mines the Communities, Regions, Activities, Topics, and Sentiments. Then an approximate learning algorithm was devised based on collapsed Gibbs sampling that estimate the model parameters of CRATS. This model was utilized in the text classification and venue recommendation. By using five latent variables the time consumption is high.

III. Sentiment Variation Tracking

In this section, the proposed hybridized Collapsed Gibbs Sampling (CGS) and Collapsed Variational Bayes (CVB) for sentiment variation tracking is described in detail. In Latent Dirichlet Allocation (LDA) based interpretation of sentiment variation model, Gibbs sampling [21] is used as approximate

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

inference methods. The major issue of Gibbs sampling is only quite effective in avoiding local optima. Moreover, the RCB-LDA [9] has inference problem. So in the proposed work, a hybridized Collapsed Gibbs Sampling (CGS) and Collapsed Variational Bayes (CVB) (HCGSCVB-LDA) model are introduced to overcome the above mentioned issues. The improved sampling method is used in both existing FB-LDA model and RCB-LDA models. The FB-HCGSCVB-LDA filters the background topics and obtains the foreground topics. From the foreground topics the more representative tweets are extracted and ranked using RCB-HCGSCVB-LDA model. Thus the performances of sentiment analysis and sentiment variations interpreting methods are improved for better decision making. Fig. 1 shows the process of twitter sentiment analysis performed in this work. It is notable that the topics filtering process provides the real improvement in the final output.

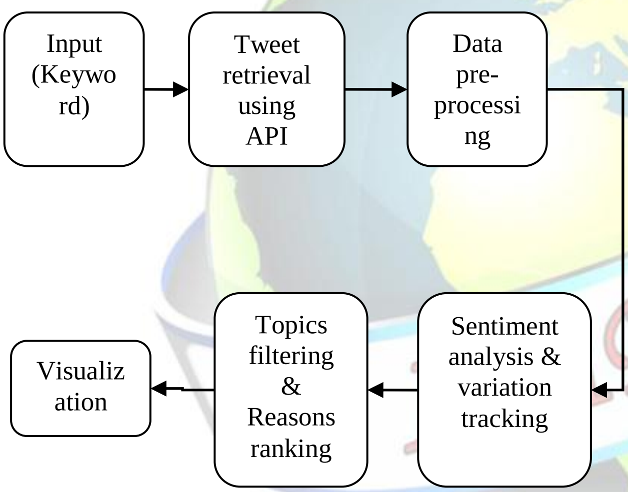

Fig. 1. Sentiment analysis & Variation tracking

A. Collapsed Gibbs Sampling (CGS)

The collapsed Gibbs Sampling algorithm allows us to compute the joint distribution $P(\vec{w},\vec{z})$ by integrating out θ^d and $\emptyset^z$ and building a Markov Chain whose transition distribution is written as:

$$P(z_x = k|\vec{z}_{-x},\vec{w}) \propto \frac{n_{k,-x}^{(w_x)}+\beta}{\sum_{w\in V} n_{k,-x}^{(w_x)}+|V|\beta}\cdot\frac{n_{k,-x}^{(d)}+\alpha}{\sum_{k=1}^{K} n_{k,-x}^{(d)}+K\alpha} \quad (1)$$

Where k is the index of latent topic , K is the total number of topic classes, $n_{k,-x}^{(.)}$ is a count that doesn't include the current assignment of z_x, and V is the vocabulary used.Given the Markov chain, one can the approximate the θ and $\emptyset$ by:

$$\emptyset_k^{(w)} = \frac{n_k^{(w)}+\beta}{\sum_{w\in V} n_k^{(w)}+|V|\beta} \quad (2)$$

$$\theta_k^{(d)} = \frac{n_k^{(d)}+\alpha}{\sum_{k=1}^{K} n_k^{(d)}+K\alpha} \quad (3)$$

The central limit theorem tells us that sums of random variables tend to concentrate and behave like a normal distribution under certain conditions. Moreover, the variance/covariance of the predictive distribution is expected to scale with $1/n$, where n is the number of data-cases that contribute to the sum. The variational approximations work well for large counts. The CVB is derived from the CGS named as hybridized CGS and CVB which can be viewed as tunable tradeoff between bias, variance and computational efficiency.

B. Hybridized Collapsed Gibbs Sampling (HCGS) And Collapsed Variational Bayes (CVB)

The proposed inference algorithm is developed by combining the advantages of both standard Collapsed Variational Bayes and Collapsed Gibbs Sampling. There are two ways to deal with the parameters in an exact fashion, the first is to marginalize them out of the joint distribution and to start from 1, and the second one is to explicitly model the posterior of θ, $\emptyset$ given z and x without any assumptions on its form. But in the proposed model only one assumption is made, i.e. the latent variables z are mutually independent.

$$\hat{q}(z,\theta,\emptyset) = \hat{q}(\theta,\emptyset|z)\prod_{i,j}\hat{q}(z_{ij}|\hat{\gamma}_{ij}) \quad (4)$$

Where $\hat{q}(z_{ij}|\hat{\gamma}_{ij})$ is multinomial *distribution* with $\hat{\gamma}_{ij}$ parameters. The key computational challenge for these models is inference, namely estimating the posterior distribution over both parameters and hidden variables; and ultimately estimating predictive probabilities and the marginal log-likelihood (or evidence). In the hybridized CGS and CVB the evidence only depends on the count arrays (N_{om} & N_{my}), N_m which are sums of assignment variable $N_{omy} = \sum_x W_{xo} H_{xm} T_{xy}$ in which H is only random. Here $n_x = o$ is the word type and $t_x = y$ is the tweet label.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

The Rao-Blackwellised estimate is a result which characterizes the transformation of an arbitrarily crude estimator into an estimator that is optimal by the mean-squared-error criterion or any of a variety of similar criteria. Rao-Blackwellised estimate of the predictive distribution is a function of these counts, which is defined as follows:

$$p(w^* = o|\{n_x, h_x, t_x\}, t^* = y) = \sum_m \frac{N_{om} + \varphi N_{my} + \alpha_m}{N_m + O\varphi N_y + \sum_m \alpha_m} \quad (5)$$

$Where N_{om} = \sum_y N_{omy}$ and $N_m = \sum_y N_{omy}$
$W_{xo} = I[n_x = o]$, $H_{xm} = I[t_x = y]$, $T_{xy} = I[h_x = m]$ are the indicator variables.

For the better variation approximation for large counts by using hybridized CGS and CVB where the dataset is split into two subsets and for one subset P^{CVB} variational Bayes approximation is applied and for another subset P^{CGS} collapsed Gibbs sampling is applied as follows:

$$P^{CVB} = \{i|\widehat{N}_{n_x, t_x} > s\} \quad (6)$$

$$P^{CGS} = \{i|\widehat{N}_{n_x, t_x} \leq s\} \quad (7)$$

In this experiment, the value of s is assigned to 1. $\widehat{N}_{n_x, t_x}$ denotes the count values. The evidence of the collapsed distribution under these assumptions defined as:

$$\varepsilon = H(P^{CVB}) + H(P^{CGS}) + \sum_{w \in V} n_{k,-x}^{(w_x)} . \sum_{k=1}^{K} n_{k,-x}^{(d)} + \alpha\beta \quad (8)$$

The variational update for $\hat{\gamma}_{ij}$ is computed and subsequently samples are drawn from it using the equivalent function Eq. It is defined as follows:

$$\hat{\gamma}_{ij} \propto \left(\alpha + Eq\left[n_{k,-x}^{(w_x)}\right]\right)\left(\beta + Eq\left[n_{k,-x}^{(d)}\right]\right)\left(K\beta + Eq\left[n_k^{(w)}\right]\right)^{-1} \quad (9)$$

The hybridized CGS and CVB is given as:

$$p(h_x^{CG} = M|h_{-x}^{CG} w, t) \propto \frac{N_{om}^{-x,sam} + \varphi}{N_m^{-x,sam} + O\varphi}(N_{ym}^{-x,sam} + \alpha_m \beta) \quad (10)$$

These update convergence in expectation and stochastically maximize the expression for the bound on the evidence. Infinitely many samples can be drawn from $\hat{\gamma}_{ij}$ to guarantee convergence. This sampling method is used in both FB-LDA and RCB-LDA to improve the convergence rate and reduce the inference problem in Gibbs sampling of LDA and it is named as Foreground and Background Hybridized CGS and CVB (FB-HCGS-CVB-LDA) and Reason Candidate and Background- Hybridized CGS and CVB- LDA respectively.

C. FB-HCGS-CVB-LDA and RCB-HCGS-CVB-LDA

The foreground and Background LDA (FB-LDA) distinguish the foreground topics from the background topics. The foreground topics reveal the possible reasons of sentiment variations in the form of word distribution. Based on the variation period of time foreground and background topics are considered. Only tweets corresponding to foreground a topic that defines emerging topics will be used to build foreground topics. So the background topics need to be filtered out from the foreground topics based on the procedure of FB-LDA. The exact inference for FB-LDA is controlled by using a hybridized Collapsed Gibbs Sampling (CGS) and Collapsed Variational Bayes (CVB) (HCGSCVB) which also improves the convergence rate. The tweets corresponding to foreground topic will be used to build the foreground topic. So background topics are filtered out from foreground topics based on the time. After the filtering process the reason candidates are found out by finding the most relevant tweet for each foreground topics learnt from FB-HCGS-CVB-LDA using the following measure:

$$Relevance(t, k_f) = \sum_{x \in t} \rho_f^{k_f, x} \quad (11)$$

In equation 11, $\rho_f^{k_f,x}$ denotes the word distribution for the foreground topic k_f and x denotes the index of each non repetitive word in tweet t. The generative process of RCB-HCGSCVB-LDA is similar to FB-HCGS-CVB-LDA. It creates the reason candidate set and background tweet sets as the standard LDA. The main difference of RCB-HCGSCVB-LDA from FB-HCGS-CVB-LDA is each word in the foreground tweets set can select a topic from alternative topic distributions. The mapping from a foreground tweet t to any background or reason candidate topic can be controlled by user defined parameter. Because of the space limit some generative process of RCB-HCGSCVB-LDA is eliminated which are similar to those in FB-HCGS-CVB-LDA. Thus RCB-HCGSCVB-LDA ranks the candidates by assigning each tweet in the foreground tweets set to one of them or the background. Table 1 and table 2 show FB-HCGS-CVB-LDA algorithm and RCB-HCGS-CVB-LDA algorithm respectively.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

TABLE I
FB-HCGS-CVB-LDA Algorithm

Algorithm 1: FB-HCGS-CVB-LDA

1. Select a word distribution based on HCGSCVB-LDA (using equation 10) for each foreground topic
2. Select a word distribution based on HCGSCVB-LDA (using equation 10) for each background topic
3. for (i=0;i<BT;i++)// BT represents the number of tweets in the background data
4. Choose a topic distribution
5. for (j=0;j<WT; i++) //WT represents the number of words in the tweet
6. Select a topic
7. Select a word
8. end for
9. end for
10. for (m=0; m<FT; m++) //FT represents the number of tweets in the foreground data
11. Choose a type decision distribution
12. for (n=0; n< WT; n++)
13. Select a type association set
14. if type is 0
15. Select a foreground distribution
16. Select a topic
17. Select a word
18. else
19. Select a background topic distribution
20. Select a topic
21. Select a word
22. end if
23. end for
24. end for

TABLE III
RCB-HCGS-CVB-LDA Algorithm

Algorithm 2 RCB-HCGS-CVB-LDA

1. for (s=0;s< FT;s++)
2. Selecta type decision distribution
3. Select a candidate association distribution
4. for (m=0; m<T; m++)// T represents the number of words in the tweets
5. Select a type association set
6. if type is 0
7. Select a Candidate
8. Select a topic
9. Select a word
10. else
11. Select a topic distribution
12. Select a topic
13. Select a word
14. end if
15. end for
16. end for

IV. Experimental Result

A. *Data Collection and Pre-processing*

For experimental purposes Twitter dataset has been considered for sentiment analysis process. Dataset are collected automatically using Twitter API and it span around 4000 tweets collected from 11.09.2016 to 20.04.2017. In this work, two key targets to test the proposed methods were selected, which are “NarendraModi” and “Apple”. This dataset has been preprocessed using unigram and bigram models in addition to removal of URLs and the non-English tweets. 200 tweets about the interested targets are selected from the collected dataset and used for training our methods and 428 tweets used for testing.

B. *Public Sentiment Analysis and Variations Tracking*

The sentiments are analyzed by combining both improved TwitterSentiment and improved SentiStrength. The final decision of the tweet label class is decided using the similar strategies as proposed by Tan et al [9] in the following steps:

- If improved TwitterSentiment and improved SentiStrength tools makes the same judgment adopt the judgment
- If the judgment of one tool is neutral and another tool is non-neutral then adopt the non-neutral judgment
- If the judgment of two tools are conflict each other, then adopt the judgment of improved SentiStrength if the FinalScore is greater than 1 otherwise adopt the judgment of improved TwitterSentiment tool.

In our experimental work the sentiment variations for the selected targets have been tracked using of

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

combining TwitterSentiment and sentiStrength [9], and Improved TwitterSentiment and Improved SentiStrength[10]. Fig. 2. shows the sentiment variation tracking of positive and negative tweets about “NarendraModi” labeled by TwitterSentiment and SentiStrength. Fig. 3. shows the sentiment variations of positive and negative tweets about “NarendraModi” labeled by Improved TwitterSentiment and Improved SentiStrength method.

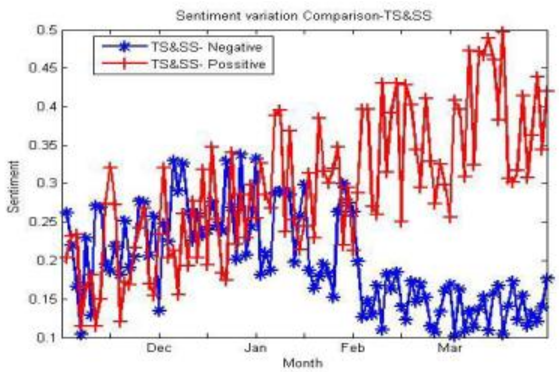

Fig.2. Sentiment variation for positive and negative terms labeled by TS&SS

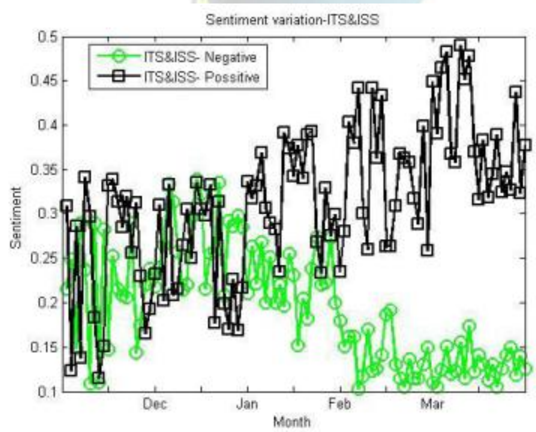

Fig. 3. Sentiment variation for positive and negative terms labeled by ITS&ISS

C. *Foreground Topics Filtering using FB-HCGS-CVB LDA based Model*

The idea of integrating out parameters before applying variational inference has been independently utilized previously. Unfortunately, because they worked in the context of general conjugate exponential families, the approach cannot be made generally computationally useful. Nevertheless, the insights of CVB can be applied to a wider class of discrete graphical models beyond LDA. Specific examples include various extensions of LDA Hidden Markov models with discrete outputs, and mixed-membership models with Dirichlet distributed mixture coefficients. These models all have the property that they consist of discrete random variables with Dirichlet priors on the parameters, which is the property allowing us to use the Gaussian approximation. Below is the top 10 foreground topics’ words extracted using FB-HCGS-CVB LDA-based model.

1. *amp modisc up ac wait obituary*
2. *lutyensprog resists devgan questioned #failedyears anointment*
3. *finally handsets made attention discrimination itnasaxena*
4. *anti islam co @narendramodi arrested asks election*
5. *sebi shah muslimsdhruv worldwide agnst launch*
6. *modi intelligence made amp anti*
7. *fali old hyderabad amp progtubelights @bainjal*
8. *under @narendramodi stand very age ias rapidly*
9. *fr demonetization court justice same seeing lost*
10. *finally handsets made attention discrimination itnasaxena*

It can be seen that the foreground topics extracted almost clear understanding such that providing proof for the proposed theory of sentiment analysis.

D. *Representative Reason Candidates’ Extraction and Ranking Using RCB-HCGS-CVB-LDA*

Representative reason candidates extracted and ranked using RCB-HCGS-CVB LDA-based model. Fig. 4 shows the top six negative tweets about “Narendra Modi” which are ranked according to their importance.

Count	Reasons
12	ken roth @kenrothjewishyahoodhihunoodmodi amp yahoodroth made each
12	today's bjp rally amit shah modigovt achievement vision mcd fake propaganda agnst @aamaadmiparty
12	new york times writes damning editorial modi'schocieadityanath up cm indian government reacts
11	withincmodi ran rat promise giving loan waiver farmers india believe #evmfraud happen
11	can't tweet anything except anti modi anti bjp can't tweet welcome back riots owaisi
10	seems yogi's anti romeo squad modi's demonetization unpopular news rooms lutyens living rooms very pop

Fig. 4. Representative reason candidates

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

The proposed algorithm has the same flavour as stochastic EM algorithms where the E-step is replaced with a sample from the posterior distribution. Similarly the proposed algorithm forms a Markov chain with a unique equilibrium distribution, so convergence is guaranteed.

E. *Performance Evaluation*

The experiments have been conducted using four methods for performing sentiment analysis and have been compared in terms of accuracy, precision, recall and in order to evaluate the performance of each method. The comparisons are made between the FB- HCGS-CVB-LDA and RCB-HCGS-CVB-LDA with combinations made with TwitterSentiment (TS) and SentiStrength (SS) and also with Improved TwitterSentiment (ITS) and Improved SentiStrength (ISS). The comparison of the performance of the models prove that the model combining FB-RCB-HCGSCVB-LDA with ITS & ISS has better performance than other comparative models.

The proposed model presents an improvement over existing models in two ways. Firstly, a more sophisticated variational approximation is used that can infer posterior distributions over the higher level variables in the model. Secondly, a more sophisticated LDA based model with an infinite number of topics is used, and allow the model to find an appropriate number of topics automatically. These two advances are coupled to provide the more sophisticated variational approximation.

1.) *Accuracy*

Accuracy is the measure of correctly labeled sentiments in all instances. It can be calculated by:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

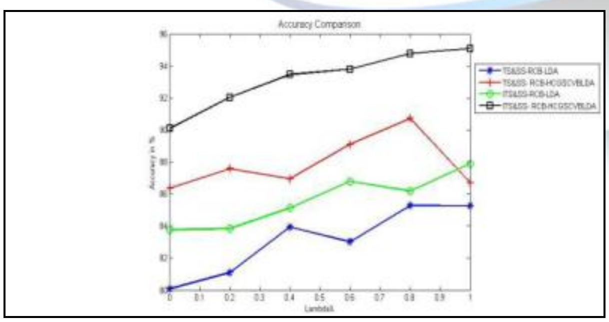

Fig. 5. Comparison of accuracy for twitter dataset

Fig. 5 shows the comparison of accuracy between TwitterSentiment (TS) and SentiStrength (SS) with Foreground and Background LDA with Reason Candidate and Background LDA(TS&SS-FB-RCB-LDA), TS and SS with FB-RCB with Hybridized Collapsed Gibbs Sampling LDA (TS&SS-FB-RCB-HCGSCVBLDA), Improved TS and Improved SS with FB-RCB-LDA (ITS&ISS-FB-RCB-LDA) and Improved TS and Improved SS with FB-RCB-HCGSCVBLDA (ITS&ISS- FB-RCB-HCGSCVBLDA) for our dataset. X axis represents λ value and Y axis represents the accuracy value in %. From the Fig. 5, it is seen that the ITS&ISS-FB-RCB-HCGSCVB-LDA has high accuracy value than the other methods.

2.) *Precision vs. Recall*

Precision value is evaluated according to the relevant information at true positive prediction, false positive.

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

The Recall value is evaluated according to true positive prediction, false negative.

$$Recall = \frac{TP}{TP + FN}$$

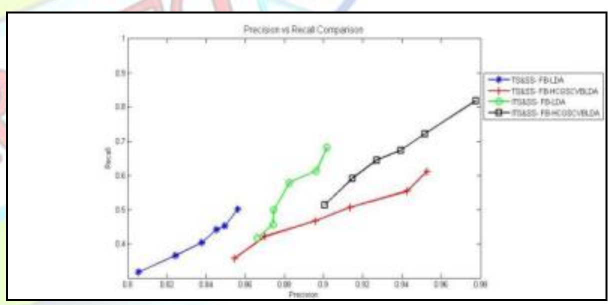

Fig. 6. Comparison of precision for twitter dataset

Fig 6. shows the comparison of precision vs. recall between Twitter Sentiment (TS) and SentiStrength (SS) with Foreground and Background LDA with Reason Candidate and Background LDA (TS&SS-FB-RCB-LDA), TS and SS with FB-RCB with Hybridized Collapsed Gibbs Sampling LDA (TS&SS-FB-RCB-HCGSCVBLDA), Improved TS and Improved SS with FB-RCB-LDA (ITS&ISS-FB-RCB-LDA) and Improved TS and Improved SS with FB-RCB-HCGSCVBLDA (ITS&ISS- FB-RCB-HCGSCVBLDA) for Twitter dataset. X axis represents precision value and Y axis represents the recall value. From the Fig. 6, it is seen that the

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 9, September 2017

ITS&ISS-FB-RCB-HCGSCVB-LDA has high recall value than the other methods.

V. CONCLUSION

This article has been aimed at deriving an improved model of Twitter sentiment analysis which can be utilized for all social networking. The major concern was the inference problem and the slow convergence rate. The proposed sentiment analysis model combining the advantages of both FB-LDA and RCB-LDA using Hybrid CGS-CVB named as FB-RCB-HCGSCVB-LDA when implemented with ITS&ISS provided high values of accuracy and higher ratio of precision- recall. This proves the efficiency of the proposed model. On deep analyzing the proposed models, it is found that adding additional measures for evaluation can further improve the sentiment variation tracking accuracy. Similarly, the detection of optimal threshold automatically can improve the final results significantly. These concepts will be examined in the future researches.

REFERENCES

[1]. B. Liu, "Sentiment Analysis And Subjectivity," In *Handbook Of Natural Language Processing, Second Edition*," Pp. 627-666, 2010.

[2]. B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*," *5*(1), 1-167, 2012.

[3]. S. Sun, C. Luo, And J. Chen, "A Review Of Natural Language Processing Techniques For Opinion Mining System," *Information Fusion*, *36*, 10-25, 2107.

[4]. J. Jiang, "Information Extraction from Text," In *Mining Text Data,"* Pp. 11-41, 2012.

[5]. R. Xia, F. Xu, J. Yu, Y. Qi, and E. Cambria, "Polarity Shift Detection, Elimination and Ensemble: A Three-Stage Model for Document-Level Sentiment Analysis," *Information Processing & Management*, *52*(1), 36-45, 2016.

[6]. D. Tang, B. Qin, F. Wei, L. Dong, T. Liu, and M. Zhou, "A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification," *Ieee/Acm Transactions on Audio, Speech, and Language Processing*, *23*(11), 1750-1761, 2015.

[7]. X. Xu, X. Cheng, S. Tan, Y. Liu, And H. Shen, "Aspect-Level Opinion Mining Of Online Customer Reviews," *China Communications*, *10*(3), 25-41, 2013.

[8]. S. Shi, D. Jin, And G. Tiong-Thye, "Real-Time Public Mood Tracking Of Chinese Microblog Streams With Complex Event Processing," *Ieee Access*, *5*, 421-431, 2017.

[9]. S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, and X. He, "Interpreting The Public Sentiment Variations On Twitter," *Ieee Transactions On Knowledge And Data Engineering*, *26*(5), 1158-1170, 2014.

[10]. O. Ghaleb, And A. Vijendran, "An Enhancement Of Public Sentiment Analysis In Social Networking By Improving Sentiment Analysis Tools," *Accepted In Journal Of Theoretical And Applied Of Information Technology*, Aug. 2107

[11]. Y. Li, X. Zhou, Y. Sun, And H. Zhang, "Design And Implementation Of Weibo Sentiment Analysis Based On Lda And Dependency Parsing," *China Communications*, *13*(11), 91-105, 2016.

[12]. J, Liang, P. Liu, J. Tan, and S. Bai, "Sentiment Classification Based On As-Lda Model," *Procedia Computer Science*, *31*, 511-516, 2014.

[13]. M. Shams, And A. Baraani-Dastjerdi, "Enriched Lda (Elda): Combination of Latent Dirichlet Allocation with Word Co-Occurrence Analysis for Aspect Extraction," *Expert Systems with Applications*, *80*, 136-146, 2017.

[14]. M. Hoffman, F.R Bach, And D.M Blei, "Online Learning For Latent Dirichlet Allocation," In *Advances In Neural Information Processing Systems*, Pp. 856-864, 2010.

[15]. Z. Hai, G. Cong, K. Chang, P. Cheng, And C. Miao, "Analyzing Sentiments In One Go: A Supervised Joint Topic Modelling Approach," *Ieee Transactions On Knowledge And Data Engineering*, 2017.

[16]. Y. Liu, J. Wang, And Y. Jiang, "Pt-Lda: A Latent Variable Model To Predict Personality Traits Of Social Network Users," *Neurocomputing*, *210*, 155-163, 2017.

[17]. M. Shams, A. Shakery, And H. Faili, "A Non-Parametric Lda-Based Induction Method For Sentiment Analysis," In *Artificial Intelligence And Signal Processing (Aisp), 2012 16th Csi International Symposium On*, Pp. 216-221, May. 2012. Ieee.

[18]. P. Chikersal, S. Poria, And E. Cambria, "Sentu: Sentiment Analysis Of Tweets By Combining A Rule-Based Classifier With Supervised Learning," In *Proceedings Of The International Workshop On Semantic Evaluation, Semeval*, Pp. 647-651, Jun. 2015.

[19]. A. Onan, S. Korukoglu, And H. Bulut, "Lda Based Topic Modeling In Text Sentiment Classification: An Empirical Analysis," In *International Journal Of Computational Linguistics And Applications, 7(1),* 110-119, 2016.

[20]. J.D. Zhang, And C.Y Chow, "Crats: An Lda-Based Model for Jointly Mining Latent Communities, Regions, Activities, Topics, and Sentiments from Geosocial Network Data," *Ieee Transactions On Knowledge And Data Engineering*, *28*(11), 2895-2909, 2016.

[21]. W.M. Darling, "A Theoretical and Practical Implementation Tutorial on Topic Modelling and Gibbs Sampling," In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.