ISSN: 2321-7782 (Online)
Volume 2, Issue 3, March 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Paper / Case Study
Available online at: www.ijarcsms.com

Affective Text based Emotion Mining in Social Media

D. Jennifer[1]
Assistant Professor
Department of Computer Science and Engineering
Apollo Engineering College
Chennai, Tamil Nadu – India

Saranya. G[2]
PG Scholar
Department of Computer Science and Engineering
Apollo Engineering College
Chennai, Tamil Nadu – India

Abstract: ***Affective text based mining of social emotion deals with new aspect for categorizing the document based on the emotions such as victory, sympathy, love etc In order to predict the emotion contained in content a joint emotion–topic model is proposed by augmenting Latent Drichlet Allocation with an additional layer for emotion modeling. Using this it first generates a set of latent topics from emotions, followed by generating affective terms from each topic. It first generates an emotion from a document-specific emotional distribution, and then generates a latent topic from a Multinominal distribution conditioned on emotions. The proposed model utilizes the complementary advantages of both emotion-term model and topic model and also it includes more websites for creating a large vocabulary. Emotion-topic model allows associating the terms and emotions via topics which is more flexible and has better modeling capability. For each emotion, a meaningful latent topic can be generated effectively and also based on emotions, song recommendation will be available for user with advantage of user uploading and enjoying their own choices also.***

Keywords: Emotion term model, Latent Drichlet Allocation, Emotion Topic model.

I. INTRODUCTION

Mining frequent patterns is probably one of the most important concepts in data mining. A lot of other data mining tasks and theories stem from this concept. It should be the beginning of any data mining technical training because, on one hand, it gives a very well shaped idea about what data mining is and, on the other, it is not extremely technical .Here affective text based mining allows us to infer a number of conditional probabilities for unseen documents, e.g., the probabilities of latent topics given an emotion, and that of terms given a topic. There are different methods used to deal with the affective text mining and following process such as, Emotion-Term model, term- based SVM model, topic based-SVM model and LDA model and so on.LDA model can only discover the topics from document and cannot bridge the connection between social emotions and affective text. Previous works mainly focuses on titles information, so the efficiency of these models is varying. Emotion-term model simply treats terms individually and cannot discover the contextual information within the document. Emotion-term model cannot utilize the term co occurrence information within document and cannot distinguish the general terms from the affective terms. On the other side, the traditional topic model can only discover the latest topics underlying the document set and cannot bridge the connections between social emotions and affective texts.

II. RELATED WORK

This section reviews some of the related work on affective text mining and topic modeling, followed by discussing their connections and differences with the proposed model.

A. Affective Text Mining:

To explore the connection between social emotions and affective terms a task named SentiWordnet is considered."A Pub-Licly Available Lexical Resource for Opinion Mining," It involves Opinion about product, political candidate from that social website user. It mainly collects information for Opinion from text and that will be based o n Text SO polarity i.e subjective and

Objective polarity and also Text PN polarity for positive and negative categorization of a opinion, from that Strength of PN polarity is calculated on basis of Score assignment to each opinion with values 0.0 to 1.0 from that total average value, weight age for opinion is being calculated.

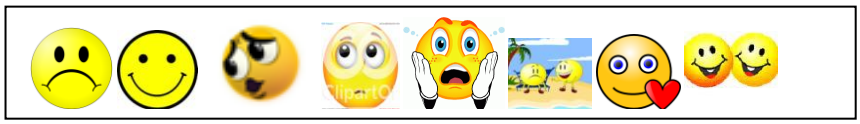


Fig 1. Emotions shown with happy & sad, sympathy & worry, surprise & devotional, love & friendship

Existing approaches will not consider relationship across word. So the emotions and terms were not linked in previous work and there will be only minimum likelihood of estimation of emotions then with the help of proposed model, we are able to visualize the emotion assignments at the term level.

B. Topic Based Analysis:

LDA has been extended to more advanced application domains with additional sampling steps even though there will be several techniques available to predict topics but the key difference lies in different sampling distributions. Their author variable is chosen uniformly from a set of authors while emotion variable is sampled from multinomial distributions by the emotions contributed by web users. LDA is extended with a different set of information, i.e., social emotions contributed by online users, in the latent topics modeling process.

C. Used Method:Social Affective Text Mining:

An online text collection D is associated with a vocabulary W, and a set of predefined emotions E. we are comparing the extracted and optimized content with the already founded latent topics that relating to each emotion. Based on the result emotions were predicted for particular represented content. Based on the user emotion request the categorized content will be displayed. To model the connections between words and emotions, and to improve the performance of its related tasks such as emotion prediction both the emotion-term and emotion-topic model can be applied to emotion prediction by estimating the probability we evaluate their prediction performance here. Whenever the user asks for any emotion in particular, it will display the contents which are all related to that user requested emotion. It has the advancement in contextual music recommendation. It not only displays the content based on the emotion, but also the songs related to that particular emotion we give. The user will upload the songs and mean while admin can also perform. It is also used in blog sites. For blogs also the emotion can be predicted.

For identifying the type of a document it refers will be based on the maximum occurred terms that were in a same group for that this process is considered

i) Term Identification Model: Naive Bayes method is used by assuming words are independently generated from social emotion labels. It generates each word wi of document d in two sampling steps, i.e., sample an emotion ei according to the emotion frequency count d, and sample a word wi given the emotion under the conditional probability P (w|e). The model parameters can be learned by maximum likelihood estimation. It can be formally derived based on the word and emotion frequency counts.

ii) Topic Identification Model: LDA addresses the over fitting problem faced by other models like pLSI by introducing a Dirichlet prior over topics and words. Although LDA cannot find connection between emotion and terms hence a more effective concept called Gibbs sampling can be used to bridge the gap between emotions and terms.

iii)Emotion Topic Model: Topic model utilizes the contextual information within the documents, it fails to utilize the emotional distribution to guide the topic generation.Topic model will categorise a document based on referred terms and then

that topic related contents will be stored in specific category. For the user the topic related content will be displayed to user from this categorized topic related contents.

III. ARCHITECTURE OF AFFECTIVE TEXT BASED MINING

Server will extract content from given URL and then these contents were optimized i.e. unwanted words will be removed. Based on categorized vocabulary the retrieved contents were stored separately in specified category in a database. For a match based on user request it will be displayed to the user.

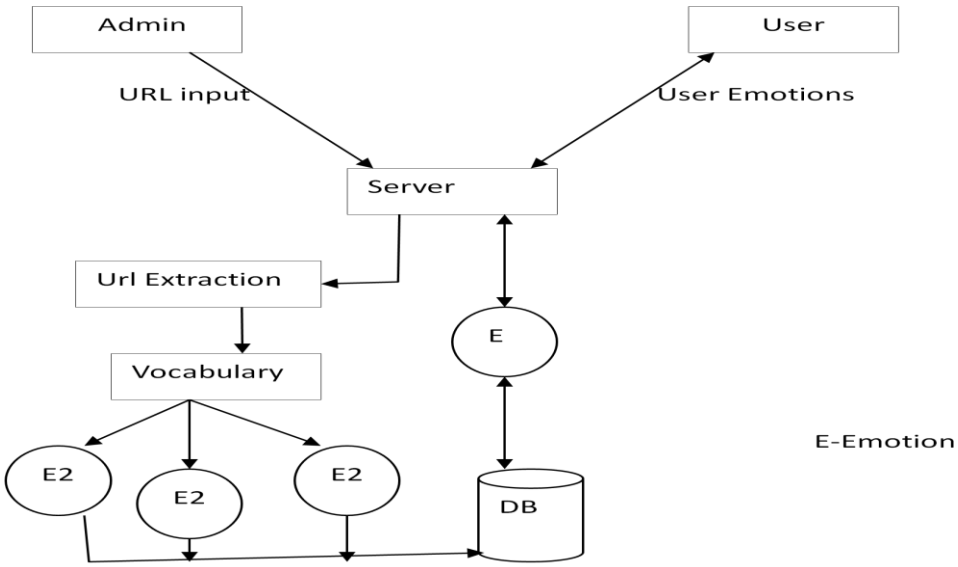


Fig 2. Affective Text Based Mining Architectur

IV. PROPOSED SCHEME

The overall process of the proposed system is depicted in Fig 4 which involves Latent Topics Generation and processing for creating a vocabulary and then Extraction and Optimization Processes to remove unwanted words which extends to the process of Social Affective Text Mining to categorise for a specific emotion which gives content to do Data based Emotion Prediction & recommendation of songs.

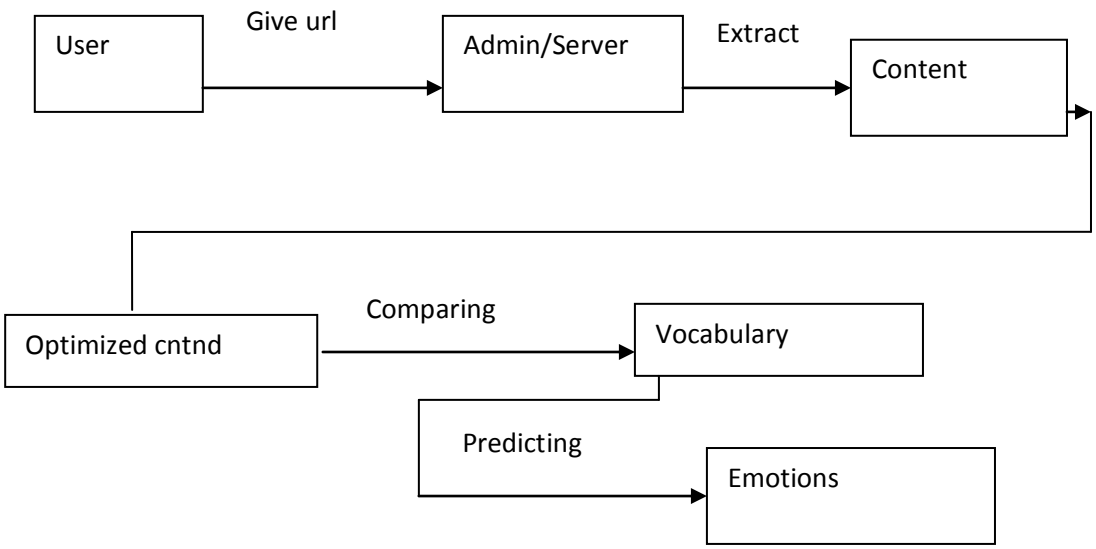


Fig 3. Affective Text Based Mining Architecture

A. Lda Algorithm

Latent Dirichlet allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

1. A word w ∈ {1, . . . , V } is the most basic unit of discrete data. For cleaner notation, w is a V - dimensional unit-based vector. If w takes on the ith element in the vocabulary, then $w^i = 1$ and $w^j = 0$ for all j 6= i.
2. A document is a sequence of N words denoted by $w = (w_1, w_2, \ldots, w_N)$, where w_n is the nth word in the sequence.
3. A corpus is a collection of M documents denoted by $D = \{w_1, w_2, \ldots, w_M\}$.
4. A topic z ∈ {1, . . . , K} is a probability distribution over the vocabulary of V words. Topics model particular groups of words that frequently occur together in documents, and thus can be interpreted as "subjects." As we will discuss in section 2.3, the generative process imagines that each word within a document is generated by its own topic, and so $z = (z_1, z_2, \ldots, z_N)$ denotes the sequence of topics across all words in a document.

1)Generative Process:Since the generative process imagines each word to be generated by a different topic, the LDA model allows documents to exhibit multiple topics to different degrees. This overcomes the limitations of the mixture of unigrams model which only allows a single topic per document. We also note that the LDA model does not attempt to model the order of words within a document. It is precisely this "bag-of-words" concept that is central to the efficiency of the LDA model.

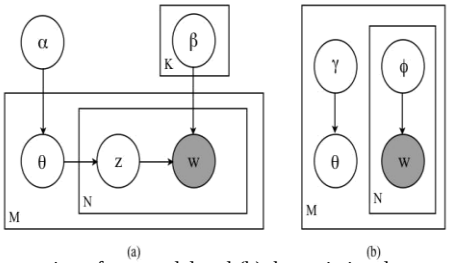
For each document indexed by m ∈ {1 . . . M} in a corpus:

1. For k = 1:::K:
   (a) $^{(k)}$ Dirichlet( )
2. For each document d 2 D:
   (a) $_d$ Dirichlet( )
   (b) For each word w_i 2 d:

z_i Discrete($_d$)

w_i Disctete($^{(zi)}$)

2) Inference and Learning: To determine the posterior distribution of the latent topic variables conditioned on the words.



(a) (b)

Fig :4 Graphical representation of our model and (b) the variational approximation (right) for the posterior distribution

B. Gibbs sampling algorithm

Gibbs Sampler allows us to to generate sample X_1; :::; X_m f(X) without requiring f(X). The motivation of Gibbs sampler is that given a multivari-ate distribution, it is simpler sample from conditional distribution than to integrate over a joint distribution. It will be shown later in this tutorial that the sequence of samples comprises a Markov Chain, and the stationary dis-tribution of the markov chain is the desired marginal distribution For example, to sample x from the joint distribution p(x) =

$p(x_1; :::; x_m)$, where there is no closed form solution for p(x), but a representation for the conditional distributions is available, using Gibbs Sampling one would perform the following

1. Randomly initialize each x_i
2. For t = 1; :::; T :

2.1 x^{t+1}_1 $p(x_1 j x^{(t)}_2; x^{(t)}_3; :::; x^{(t)}_m)$

2.2 x^{t+1}_2 $p(x_2 j x^{(t+1)}_1; x^{(t)}_3; :::; x^{(t)}_m)$

m x^{t+1}_m $p(x_m j x^{(t+1)}_1; x^{(t+1)}_2; :::; x^{(t+1)}_{m\ 1})$

This procedure is repeated a number of times until the samples begin to con-verge to what would be sampled from the true distribution

1) Algorithm Of LDA Gibbs Sampling:

Input: words w 2 documents d

Output: topic assignments z and counts $n_{d;k}$; $n_{k;w}$; and n_k begin

randomly initialize z and increment counters for each iteration do

for i = 0 ! N 1 do word w[i] topic z[i] ^{n}d;topic$^{-=1;\ n}$word;topic$^{-=1;\ n}$topic$^{-=1}$

for k = 0 ! K 1 do

$p(z = k j) = (n_{d;k} +\ _k)\ ^{nk;w+\ w}$

$n_k +$ W

end

topic sample from p(zj) z[i] topic^{n}d;topic$^{+=1;n}$word;topic$^{+=1;\ n}$topic$^{+=1}$

end end

return z, $n_{d;k}$; $n_{k;w}$; n_k

end

C. *Emotion Term Model*

It follows the Naive Bayes method by assuming words are independently generated from social emotion labels. Fig. Illustrates the generation process of the emotion-term model. It generates each word ω_i of document *d* in two sampling steps, i.e., sample an emotion e_i according to the emotion frequency count γ_d, and sample a word ω_i given the emotion under the conditional probability $P(\omega|e)$.

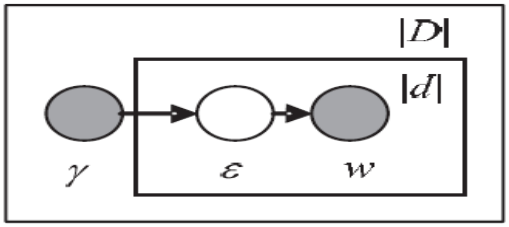


Fig 5. Emotion Term Model

The model parameters can be learned by maximum likelihood estimation. In particular, the conditional probability of a word ω given an emotion e can be estimated as follows:

$$P((\omega|e) = \frac{|(\omega, e)|}{\sum \omega` \epsilon W^{|\omega`, e|`}}$$

Where $|(\omega, e)|$ is the co occurrence count between word $\omega \in W$ and emotion $e \in E$ for all the documents. It can be formally derived based on the word and emotion frequency counts

$$|\omega, e| = S + \sum_{d \epsilon D} \delta_{d,\omega^{\gamma d,e,}}$$

where S is a small smoothing constant, for predicting emotion on a new document d, apply the Bayes theorem under the term independence assumption

$$P(e|d) = \frac{P(d|e)P(e)}{P(d)} \propto P(d|e)P(e) = P(e) \coprod_{\omega \epsilon d} P(\omega|e)^{\delta_{d,\omega}}$$

Where $P(e)$ is the a priori probability of emotion e.

D. *Topic Model*

LDA addresses the over fitting problem faced by other models like pLSI by introducing a Dirichlet prior over topics and words. Although LDA can only discover the topics from document and cannot bridge the connection between social emotion and affective text, for the ease of understanding in the following description, a simple review of LDA here.

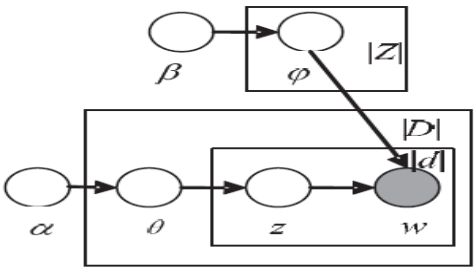


Fig 6.Topic Model

The Topic Model assumes the following generative process for each word ω_i from topic z_i in ocument di∈D

$\theta \sim Dirichlet(\alpha)$

$z_{i|}|\theta_{di} \sim$ Multi-Nomial((θ_{di})

$\varphi \sim Dirichlet(\beta)$

$\omega_i|Z_i, \varphi_{zi} \sim Multi - Nomial(\varphi_{zi})$

Where α and β are hyper parameters, specifying the Dirichlet priors on θ and φ. Here an alternative parameter estimation method, Gibbs sampling, which is more computationally efficient.

$$P(Z_{i=}j|Z_{-i}, W) \propto \frac{n_{-i,j}^{(\omega i)} + \beta \quad n_{-i,j}^{(di)} + \alpha}{n_{-i,j}^{(.)} + |W|\beta n_{-i,.}^{(di)} + |Z|\alpha`}$$

Where, n_{-i} means the count that does not include the current assignment of z_i, $n_j^{(\omega)}$ is the number of times word w has been assigned to topic j, and $n_j^{(d)}$ is the number of times a word from document d has been assigned to topic j

E. Emotion-Topic Model

Emotion-term model simply treats terms individually and cannot discover the contextual information within the document. However, intuitively, it is more sensible to associate emotions with a specific emotional event/topic instead of only a single term.

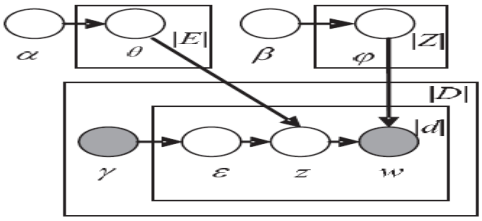


Fig 7.Emotion Topic model

As illustrated in Fig, the emotion-topic model accounts for social emotions by introducing an additional emotion generation layer to Latent Dirichlet Allocation. For each document d, this model follows a generative process for each of its words ω_i as follows:

$\theta \sim Dirichlet(\alpha)$

$\varepsilon_i \sim Multinominal(\gamma)$

$z_i|\varepsilon_i, \varphi \sim Multinominal(\theta_{\varepsilon i})$

$\varphi \sim Dirichlet(\beta)$

$\omega_i|z_i, \varphi \sim Multinominal(\varphi_{zi})$

Where α and β are hyper parameters for the Dirichlet priors on θ and φ, respectively. ε is sampled from a multinominal distribution parameterized by γ. $\varepsilon_i \epsilon E$ and $z_i \epsilon Z$ are corresponding emotion and topic assignment for word ωi. Z here means the set of topics used for emotion-topic modeling. The emotion frequency count γ is normalized and summed to 1 in this case, which allows it to serve as the parameters for the multinominal distribution. According to this generative process, the joint probability of all the random variables for a document collection is

$$P(\gamma, \varepsilon, z, w, \theta, \varphi, \alpha, \beta) = P(\theta; \alpha)P(\varphi; \beta)$$

$$P(\gamma)P(\varepsilon|\gamma)P(z|\varepsilon, \theta)P(w|z, \varphi).$$

Fr each word, estimate the posterior distribution on emotion ε and topic z based on the following conditional probabilities, which can be derived by marginalizing the above joint probabilities in

$$P(\varepsilon_i = e|\gamma, \varepsilon_{-i}, z, w; \alpha, \beta)$$

$$\propto \frac{\alpha + nz_{-i}^{e,zi}}{|Z|\alpha + \sum_z nz_{-i}^{e,z}} \times \frac{\gamma_{di,e}}{\sum e` \gamma_{di,e`}}$$

$$P(z_i = z|z_{-i}, \gamma, \varepsilon, w; \alpha, \beta)$$

$$\propto \frac{\alpha + nz_{-i}^{\varepsilon i,z}}{|Z|\alpha + \sum z` nz_{-i}^{\varepsilon i,z`}} \times \frac{\beta + n\omega_{-i}^{z,\omega i}}{|W|\beta + \sum_\omega n\omega_{-i}^{z,\omega`}}$$

Where, e and z are the candidate emotion and topic for sampling, respectively. $d_i \in D$ indicates the document from which current word ω_i is sampled. $nz^{e,z}$ is the number of times topic z has been assigned to emotion e. Similarly, $n\omega^{z,\omega}$ means the number of times a word w has been assigned to topic z. The suffix -i of nz and $n\omega$ means the count that does not include the current assignment of emotion and topic for word ω_i, respectively. In more details, we start the algorithm by randomly assigning all the words to emotions and topics. Then, repeat Gibbs sampling on each word in the document collection by applying (6) and

(7) sequentially. This sampling process is repeated for N iterations when the stop condition is met. With the sampled topics and emotions available, it is easy to estimate the distributions of ε, θ, and φ as follows.

Table1:Social Emotion Assignment Statistics.

1	Suicide,incident,accident,salvag	Sadness
	e,corpse,stress	
2	Grasp,discover,astonish,stunned	surprise
3	Annoyance,violence,irritation,m	anger
	ad,hatred	
4	Disambiguations,affinity,appreci	empathy
	ation,concord,insight,soul	
5	Cheer,action,delight,laughter,ple	amusement
	asure,enjoyment	
6	Affection,kindness,compassion,	love
	virtue,romantic	
7	Fatigue,disguist,indifference,dull	boredom
	ness,lack of interest	
8	Glowing,hot,sunny,sweating,per	warmness
	Spiring	
9	Insist,transplant,impressed,emoti	touched
	onal,grabbed	
10	Delighted,blessed,overenjoyed,	happy

V. IMPLEMENTATION DETAIL

The proposed system will show Retrieved content from a url to display html tags, from that only other tags have been removed to get original content as shown in Fig where all the categorized content from the database will get displayed with emotion based categories.

Fig 8. Viewing categorized data

VI. CONCLUSION

Social affective text mining aims to discover and model the connections between online documents and user-generated social emotions. To this end, joint emotion-topic model is proposed by augmenting Latent Dirichlet Allocation with an intermediate layer for emotion modeling. Rather than emotion term model that treats each term in the document individually and LDA topic model that only utilizes the text co occurrence information, emotion-topic model allows associating the terms and emotions via topics which is more flexible and has better modeling capability. Experimental results on an online news collection show that the model is not only effective in extracting the meaningful latent topics, but also significantly improves the performance of social emotion prediction compared with the baseline emotion-term model and multiclass SVM. The process of including more & more documents can be implemented that makes the datas availability large and for emotion categorization not only documents but also songs were also uploaded by admin and also by the user. Categorised songs which were similar to documents will be stored separately and displayed to user.

For the future work, evaluation with a larger scale of online document collections, and applying the model to other applications such as emotion-aware recommendation of advertisements can be analyzed and also instead of emotion prediction from user the process of automatic identification of emotion from user through face recognition techniques can make the process more flexible to predict emotions.

Acknowledgement

With the cooperation of my guide, I am highly indebted to Associate Prof. Dr. S. Prasanna Devi for her valuable guidance and supervision regarding my topic as well as for providing necessary information regarding review paper. I am very much thanks to Asst Prof. Ms.D. Jennifer helping me in project guidance.

References

1. C.O. Alm, D. Roth, and R. Sproat, "Emotions from Text: Machine Learning for Text-Based Emotion Prediction," Proc. Joint Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP '05), pp. 579-586, 2005.
2. Shenghua Bao,Shengliang Xu,Li Zhang, Rong Yan,Zhong Su, Dingyi Han, and Yong Yu "Mining Social Emotions from Affective Text" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 9, SEPTEMBER 2012
3. R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma., "Musicsense: Contextual Music Recommendation Using Emotional Allocation," Proc. 15th Int'l Conf. Multimedia, pp. 553-556, 2007
4. A. Esuli and F. Sebastiani, "Sentiwordnet: A Pub-Licly Available Lexical Resource for Opinion Mining," Proc. Fifth Int'l Conf. Language Resources and Evaluation (LREC '06), 2006.
5. T. Griffiths and M. Steyvers, "Finding Scientific Topics," Proc. Nat'l Academy of Sciences USA, vol. 101, pp. 5228-5235, 2004.
6. M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '04), pp. 168-177, 2004.
7. C. Strapparava and R. Mihalcea, "Learning to Identify Emotions in Text," Proc. 23rd Ann. ACM Symp. Applied Computing (SAC '08), pp. 1556-1560, 2008.
8. C. Strapparava and R. Mihalcea, "Semeval-2007 Task 14: Affective Text," Proc. Fourth Int'l Workshop Semantic Evaluations (SemEval'07), pp. 70-74, 2007.
9. C. Strapparava and A. Valitutti, "Wordnet-Affect: An Affective Extension of Wordnet," Proc. Fourth Int'l Conf. Language Resources and Evaluation (LREC '04), 2004

AUTHOR(S) PROFILE

Ms. D. Jennifer is currently working as Asst Professor in the Department of Computer science & Engineering in Apollo Engineering College, Chennai having an experience of 4.5 years. This college has been affiliated to Anna University, Tamil Nadu, and India. The M.E Degree is awarded to her by the St Peter's University during December 2012. Her area of interest includes Data Mining, Networking.

Ms Saranya. G received the B.Tech degree in Information Technology from Adhiparasakthi Engineering College of Anna University 2008, Chennai and pursuing M.E degree in Computer Science & Engineering from Apollo Engineering College of Anna University.

