ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Categorization of Web Content Based on Social Emotions Using Data Mining

P.Thilagavathi[1], P. Gayathri*[2,]

[1] Assistant Professor, Department of IT, Jerusalem College of Engineering, Chennai, Tamil Nadu, India

[2] Assistant Professor, Department of IT, Bharath University, Chennai, Tamil Nadu, India

* Corresponding Author

ABSTRACT: This paper is concerned with the problem of mining social emotions from text. Recently, with the fast development of web 2.0, more and more documents are assigned by social users with emotion labels such as happiness, sadness, and surprise. Such emotions can provide a new aspect for document categorization, and therefore help online users to select related documents based on their emotional preferences. Useful as it is, the ratio with manual emotion labels is still very tiny comparing to the huge amount of web enterprise documents. In this paper, we aim to discover the connections between social emotions and affective terms and based on which predict the social emotion from text content automatically. More specifically, we propose a joint emotion-topic model by augmenting Latent Dirichlet Allocation with an additional layer for emotion modeling. It first generates a set of latent topics from emotions, followed by generating affective terms from each topic. Experimental results on an online news collection show that the proposed model can effectively identify meaningful latent topics for each emotion. Evaluation on emotion prediction further verifies the effectiveness of the proposed model.

KEYWORDS: *Text Mining, Social Emotions, Dirichlet Allocation*

I. INTRODUCTION

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High- quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. In order to predict the emotion contained in content, we propose a joint emotion –topic model 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., which first generates an emotion from a document-specific emotional distribution, then generates a latent topic from a Multinomial distribution conditioned on emotions, The proposed emotion-topic model utilizes the complementary advantages of both emotion -term model and topic model. Emotion-topic model [7] allows associating the terms and emotions via topics which is more flexible and has better modeling capability. The model is not only effective in extracting the meaningful latent topics, but also significantly improves the performance of social emotion prediction.

A. Objective

The objective of this project is to find the connections between emotions and affective terms by categorizing the web-content, based on the emotion present in it and also predicting the emotions from text automatically.

B. Scope of project

Emotions can provide a new aspect for document categorization, and therefore help online users to select related documents based on their emotional preferences we propose a joint emotion-topic model by augmenting Latent Dirichlet Allocation [2] with an additional layer for emotion modeling. It first generates a set of latent topics from

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

emotions, followed by generating affective terms from each topic. The user-generated social emotions [1] provide a new aspect for document categorization, and they cannot only help online users select related documents based on emotional preferences, but also benefit a number of other applications such as contextual music recommendation. In this paper, we refer to the problem of discovering and mining connections between social emotions and online documents as social affective text mining [1], including predicting emotions from online documents, associating emotions with latent topics, and so on.

II. RELATED WORKS

The research in "Learning to Identify Emotions in Text [1]" by C.Strapparava and R.Mihalcea is concerned with the problem of mining social emotions from text. Recently, with the fast development of web 2.0, more and more documents are assigned by social users with emotion labels such as happiness, sadness, and surprise. Such emotions can provide a new aspect for document categorization, and therefore help online users to select related documents based on their emotional preferences. Useful as it is, the ratio with manual emotion labels is still very tiny. Comparing to the huge amount of web/enterprise documents. In this paper, we aim to discover the connections between social emotions and affective terms and based on which predict the social emotion from text content automatically.

Another work proposed by *I. Titov and R. McDonald*, "A Joint Model of Text and Aspect Ratings for Sentiment [2]"is concerned with the problem of mining social emotions from text. Recently, with the fast development of web 2.0, more and more documents are assigned by social users with emotion labels such as happiness, sadness, and surprise. Such emotions can provide a new aspect for document categorization, and therefore help online users to select related documents based on their emotional preferences. Useful as it is, the ratio with manual emotion labels is still very tiny comparing to the huge amount of web/enterprise documents [3]. In this paper, we aim to discover the connections between social emotions and affective terms and based on which predict the social emotion from text content automatically. More specifically, we propose a joint emotion-topic model by augmenting Latent Dirichlet Allocation [2] with an additional layer for emotion modeling. It first generates a set of latent topics from emotions, followed by generating affective terms from each topic. Experimental results on an online news collection show that the proposed model can effectively identify meaningful latent topics for each emotion. Evaluation on emotion prediction further verifies the effectiveness of the proposed model.

A. Existing scenario

In the existing system 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 [7] 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. The main drawbacks of the existing sytems are, the systems are found only in Chinese, low efficiency and they fail to uncover strong emotions.

III. SYSTEM DESIGN

System Architecture is the conceptual model that defines the structure, behavior, and more views of system.An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structure of the system which comprises system, components the externally visible properties of those components, the relationships between them, and provides a plan from which products can be procured, and systems developed, that will work together to implement the overall system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

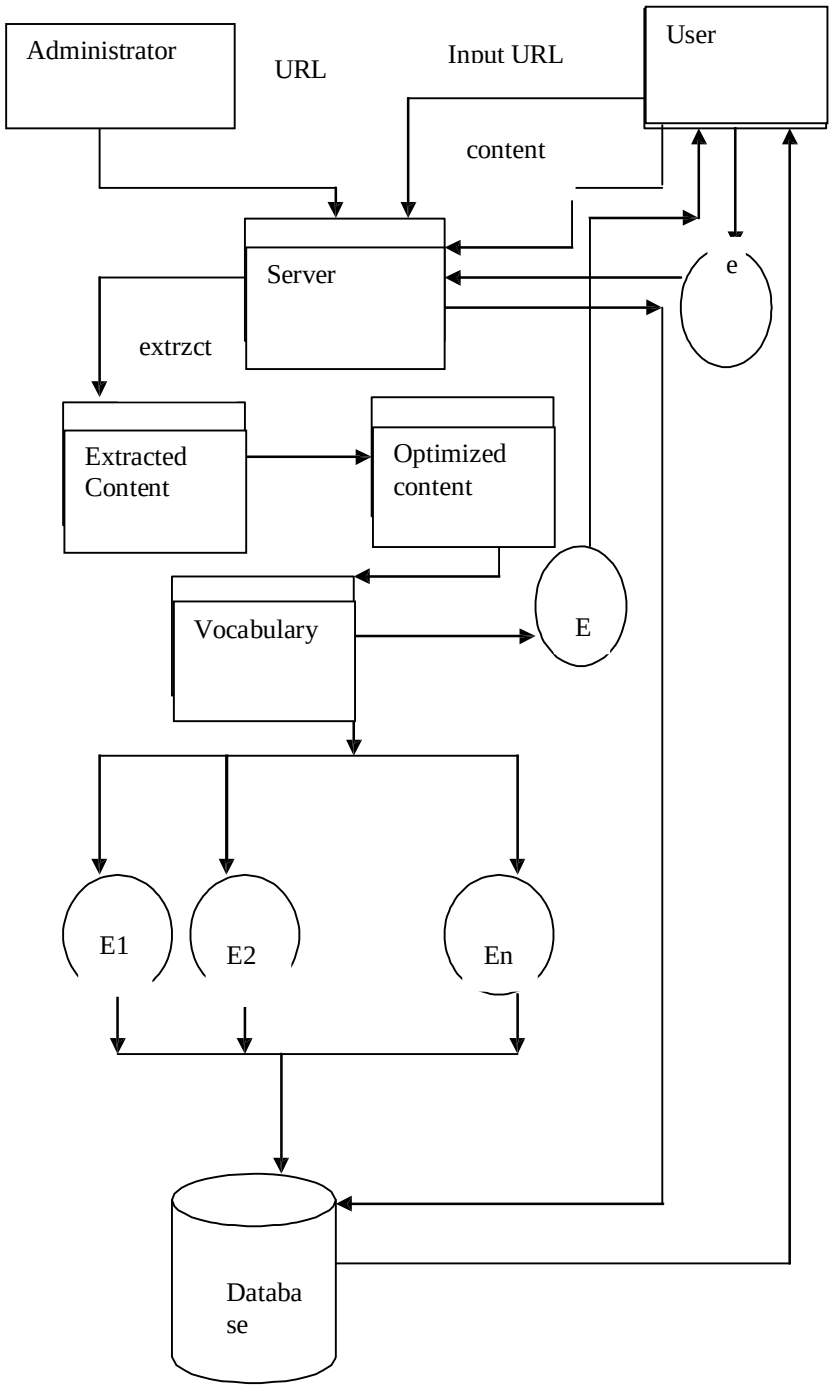

Figure 1. System Architecture

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

IV. PROPOSED SYSTEM

In the proposed system, to increase the efficiency of the processes we are using a joint Emotion topic model for social affective text mining which introduces an additional layer of emotion modeling in to Latent Drichlet Association. The proposed emotion topic model allows us to infer a number of conditional probabilities of unseen documents, the probabilities of latent topics given an emotion, and that of terms given a topic. Our objective is to accurately model the connections between words and emotions, and improve the performance of its related tasks such as emotion prediction. The emotion-topic model accounts for social emotions by introducing an additional emotion generation layer to Latent Dirichlet Allocation [2]. For each document d, this model follows a generative process for each of its words „wi" because it is intractable to perform an exact inference for the emotion-topic model, we develop an approximate inference method. The key advantage for the proposed emotion-topic model is its ability to uncover hidden topics that exhibit strong emotions[16].The outline of the proposed scheme is shown in the figure 1.The system has two processes namely admin and user.[1] First the admin enters an URL into the server. Then the URL is scanned for emotion related words by removing the html tags and non emotion related words are removed.The output from this process is known as optimized output.Then each word of the optimized output is compared with a vocabulary list that contains equivalent words for a specific emotion.Finally based on the comparison results a particular document is categorized under a particular emotion.The process is repeated for several URL"S. When the user enters a specific emotion the set of URL"s already categorized is presented to him.[3]

V. MODULES AND OPERATIONS

A. Generation and processing of latent topic

In our first module, we have to generate latent topics [6] for each emotion that we are going to consider. Each emotion we are taking and creating latent topics [5] on that particular emotion .The quantity and quality of the latent topics is more, then the efficiency of the affective text mining also more. After collecting and categorizing each latent topic [8] based on different emotions, are stored to check with the extracted content.[7]

B. Extraction of optimized text

In our second module we are going to extract the title and main body required articles. The access of the main body of articles provides the basis for modeling latent topics [4] and helps alleviate the issue of data sparseness. Segment all the words for each article.[8] Apply named entity recognition to filter out person names from the documents, because we found that few of the person names occurring in news articles bear any consistent affective meanings. After that remove all the words that represent no meaning related to any of the specified emotions and by thus optimizing the content.[9]

Algorithm

C_k: Candidate itemset of size k
L_k : frequent itemset of size k

L_1 = {frequent items};

for (k = 1; L_k != ; k++) **do begin** C_{k+1} = candidates generated from L_k; **for each** transaction t in database do increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1} = candidates in C_{k+1} with min_support

end

return $\cup_k L_k$;

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

C. *Prediction of emotion content*

In our final module an online text collection D is associated with a vocabulary W, and a set ofpredefined emotions E. In particular, each document d belongs to D consists of a number of words {wi}, wi belongs to W, and a set of emotion labels {ek}, ek belongs to E. For each emotion e, we finding the frequency count the count of each word w. [10] Here we are comparing the extracted and optimized content with the already founded latent topics that relating to each emotion. Based on the result we are finding which emotion the particular content represents. Based on the user emotion request the categorized content will display [6]. Our objective is to accurately model the connections between words and emotions, and improve the performance of its related tasks such as emotion prediction.[11]

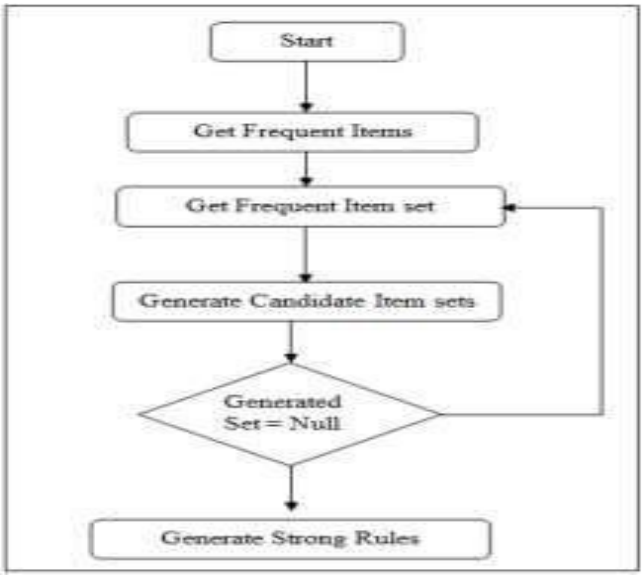

Figure 2: Prediction of Emotion Content

VI. PERFORMANCE EVALUATION

The efficiency of the existing and the proposed system are plotted against the number of URLs based on practical experiments. Thus, we find that, the proposed system turns out to be more efficient than the existing system using the following graphs.[12]

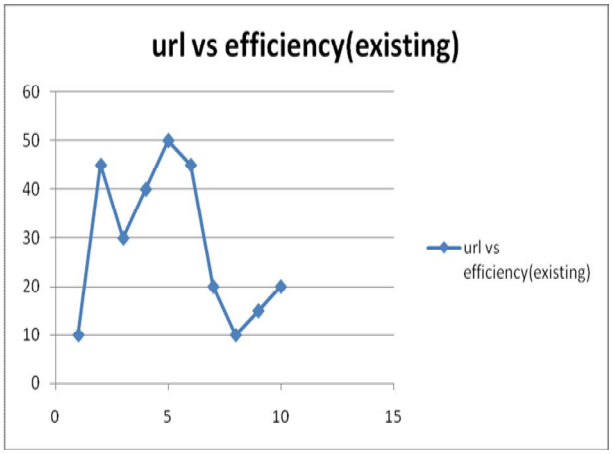

Figure 3: Existing System

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

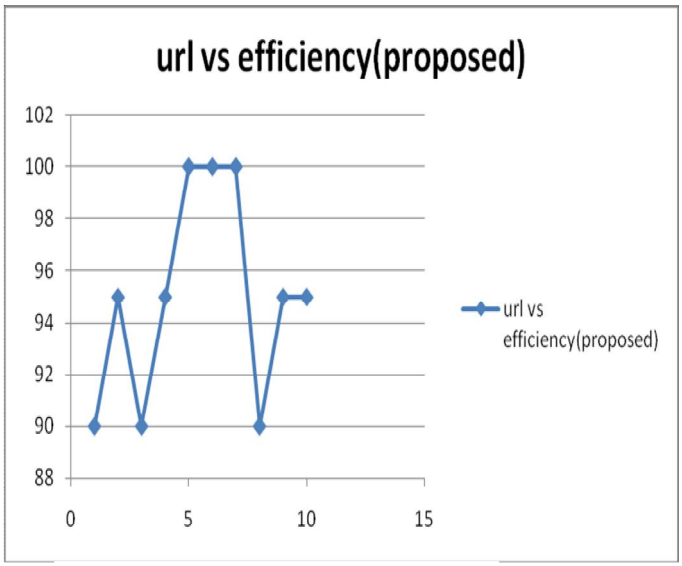

Figure 4: Proposed System

VI. CONCLUSION

Thus the proposed system predicts the emotion contained in various web documents, categorizes them and present those documents to the user based on his query. The current implementation of the project only predicts the emotion contained in web documents. The system can be extended to predict the emotion of songs based on various parameters like frequency range, bass, bgm, etc. The current system is generic in nature and it can be used in many real time applications.

REFERENCES

[1] Shenghua Bao, Li Zhang," Mining social emotions from affective text,"vol.24, no.9, sep 2012.
[2] D.M.Blei, A.Y.Ng, and M.I.Jordan, "Latent Dirichlet Allocation," J.Machine Learning Approach, vol.3, pp.993-1022, 2011.
[3] Sree Latha R., Vijayaraj R., Azhagiya Singam E.R., Chitra K., Subramanian V., "3D-QSAR and Docking Studies on the HEPT Derivatives of HIV-1 Reverse Transcriptase", Chemical Biology and Drug Design, ISSN : 1747-0285, 78(3) (2011) pp.418-426.
[4] M.Hu and B.Liu, "Mining and Summarizing Customer Reviews," Proc.10th ACM SIGKDD int˝l Conf. Knowledge Discovery and Data Mining (SIGKDD„O4),pp. 168-177,2011.
[5] W.H. Lin, E. Xing, and A. Hauptmann, "A Joint Topic and Perspective Model for Ideological Discourse," Proc. European Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD ˝08), pp. 17-32, 2008.
[6] Masthan K.M.K., Aravindha Babu N., Dash K.C., Elumalai M., "Advanced diagnostic aids in oral cancer", Asian Pacific Journal of Cancer Prevention, ISSN: 1513-7368, 13(8) (2012) pp.3573-3576.
[7] T. Griffiths and M. Steyvers, "Finding ScientificTopics," Proc. Nat˝l Academy of Sciences USA, vol.101, pp. 5228-5235, 2004. [29] P.-C.Chang, M. Galley, and C. Manning, "Optimizing Chinese Word Segmentation for Machine Translation Performance," Proc. Assoc. for Computational Linguistics (ACL) Third Workshop Statistical Machine Translation, 2008.
[8] Tamilselvi N., Dhamotharan R., Krishnamoorthy P., Shivakumar, "Anatomical studies of Indigofera aspalathoides Vahl (Fabaceae)", Journal of Chemical and Pharmaceutical Research, ISSN : 0975 – 7384 , 3(2) (2011) pp.738-746.
[9] J. Guo, S. Xu, S. Bao, and Y. Yu,"Tapping on the Potential of q&A Community by Recommending Answer Providers," Proc. ACM 17th Conf. Information and Knowledge Management (CIKM ˝08), 2008.
[10] Devi M., Jeyanthi Rebecca L., Sumathy S., "Bactericidal activity of the lactic acid bacteria Lactobacillus delbreukii", Journal of Chemical and Pharmaceutical Research, ISSN : 0975 – 7384 , 5(2) (2013) pp.176-180.
[11] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed- Membership Models of Scientific Publications," Proc. Nat˝l Academy of Sciences USA, vol. 101, pp.5220-5227, 2004.
[12] Reddy Seshadri V., Suchitra M.M., Reddy Y.M., Reddy Prabhakar E., "Beneficial and detrimental actions of free radicals: A review", Journal of Global Pharma Technology, ISSN : 0975-8542, 2(5) (2010) pp.3-11.
[13] R.M. Nallapati, A. Ahmed, E.P. Xing, and W.W. Cohen, "Joint Latent Topic Models for Text and Citations," Proc. 14th ACM SIGKDD Int˝l Conf. Knowledge Discovery and Data Mining (KDD ˝08), pp.542-550,2008.
[14]B Karthik, TVUK Kumar, A Selvaraj, Test Data Compression Architecture for Lowpower VLSI Testing, World Applied Sciences Journal 29 (8), PP 1035-1038, 2014.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

[15].M.Sundararajan .Lakshmi,"Biometric Security system using Face Recognition", Publication of International Journal of Pattern Recognition and Research. July 2009 pp. 125-134.
[16].M.Sundararajan," Optical Sensor Based Instrumentation for correlative analysis of Human ECG and Breathing Signal", Publication of International Journal of Electronics Engineering Research, Research India Publication, Volume 1 Number 4(2009). Pp 287-298.
[17]C.Lakshmi & Dr.M.Sundararajan**, "**The Chernoff Criterion Based Common Vector Method: A Novel Quadratic Subspace Classifier for Face Recognition**"** Indian Research Review, Vol.1, No.1, Dec, 2009.
[18]M.Sundararajan & P.Manikandan," Discrete wavelet features extractions for Iris recognition based biometric Security", Publication of International Journal of Electronics Engineering Research, Research India Publication, Volume 2 Number 2(2010).pp. 237-241.
[19]M.Sundararajan, C.Lakshmi & .M.Ponnavaikko, "Improved kernel common vector method for face recognition varying in background conditions", proceeding of Springer – LNCS 6026- pp.175-186 (2010).ISSN 0302-9743**.(Ref. Jor – Anne-II)**