An Efficient System for Recognizing Emotions in Text via Topic Modelling

Soumya Chandran
PG Scholar
Department of Computer Science and Engineering
K .C.G College of Technology
Chennai, India

Mr. S. Bairavel
Assistant Professor
Department of Computer Science and Engineering
K .C.G College of Technology
Chennai, India

***Abstract*— The main stay 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. Emotions can provide a new aspect for categorizing different documents, and therefore it will be useful for online users to select related documents based on their emotional preferences. In order to predict the emotion contained in a text, a joint emotion topic model by augmenting Latent Dirichlet Allocation (LDA) with an additional layer for emotion modeling is used. Using this, it first generates a set of latent topics from emotions, followed by generating emotional terms from each topic., which finally provides a specific emotion for a particular content or text from a document-specific emotional distribution. The emotion-topic model utilizes the complementary advantages of both emotion-term model and emotion-topic model. Emotion-topic model allows associating the terms and emotions via topics which is more flexible and has better modeling potential. This model notably improves the performance of social emotion prediction.**

***Key Words*— Affective terms, Emotion modelling, Emotion-term model, Latent Dirichlet Allocation(LDA).**

1. INTRODUCTION

In this 21st century, almost all the online users spend most of their time on social websites. Hence numerous social websites now provide a service that allows users to share their emotions after browsing different articles. The user-generated social emotions provide a new aspect for document categorization, and they help online users to select related documents based on their emotional preferences. Each and every article or text can evoke a pleasant or agonizing experience because of their glossologic relation to emotional concepts. However, the way how text documents affect online users' social ardour is yet to be unveiled. Automatic detection of emotions refers to the problem of recognizing and mining connections between social emotions and online documents, including prophesying emotions from online documents and associating emotions with hidden topics. A straightforward method is to manually build a dictionary of affective terms for each ardour. But, building a dictionary is not only labor engrossing, but also unable to quantize the connection strengths between affective terms and social emotions. As an alternative, Naive Bayes provides a principled way to estimate term-emotion associations using their co-occurrence counts. In this work, a joint emotion-topic model for social affective text mining, which introduces a subsidiary layer of emotion modeling into Latent Dirichlet Allocation (LDA), is used. In more details, the model follows a three-step generation process for affective terms, which first generates an ardour from a document-specific emotional distribution, then generates a hidden topic from a Multinominal distribution conditioned on ardour, and finally generates document terms from another Multinominal distribution based on latent topics. Text based emotion detection systems have gone a step beyond simple word matching by performing a semantic analysis of the text. As a complete generative model, the emotion-topic model allows us to infer a number of conditional probabilities for unseen documents, e.g., the probabilities of hidden topics given an ardour, and that of terms given a topic.

We have evaluated this model on www.thehindu.com and applied this for several articles. It has been found that this model works effectively in discovering meaningful latent topics from news documents. This model can also distinguish the topics with strong emotions from disquieting ones. For social emotion prediction, proposed model predominates emotion-term model, term-based SVM model, and topic-based SVM model substantially by 34.3, 8.10, and 15.31 percent, respectively, in terms of accuracy, which further verifies the effectiveness of this model. Fig 1.1 shows the frequency of each emotion identified in a particular document. Here we can see that the emotion happy has the highest frequency and hence conclude that this particular document contains happy content. The other emotions identified from this document are sad, surprise, sympathy, love, devotional, friendship among which the emotion 'devotional' shows the minimum frequency.

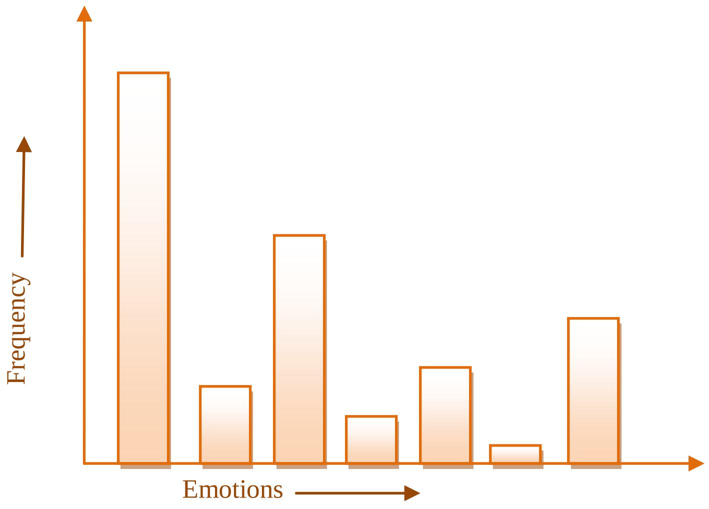


Fig 1.1 An example showing frequency of different emotions based on an emotional content. The emotions from left to right are happy, sad, surprise, sympathy, love, devotional, friendship.

This paper mainly focused on the emotion classification of different documents extracted from social web sites. Documents may sometimes contain only a few words and are often written by creative people with the intention to "provoke" emotions, and consequently to attract the readers' engrossment. These specialities make this type of text particularly suitable for use in automatic emotion recognition.

2. RELATED WORKS

The topic 'Detection of Emotions in Text' have been discussed and evaluated using different algorithms, classification methodologies etc by several authors. Carlo Strapparava and Rada Mihalcea[3] introduced a work to identify emotions in text in 2008. This paper describes experiments concerned with the automatic analysis of emotions in text. This follows a rule-based system using a linguistic approach. A first pass through the data "uncapitalizes" common words in the title of the content. The system then used the Stanford syntactic parser on the modified titles, and identifies what is being said about the main subject by exploiting the dependency graph obtained from the parser. Each word is first rated separately for each emotion and then the main subject rating is boosted. Several knowledge-based and corpus based methods are used for the automatic identification of emotions in text. Knowledge-based (WORDNET and LSA) and corpus-based (Naïve Bayes classifier) classification methods are used here. In this work, experiments for the automatic annotation of emotions in text are done. Though comparative evaluations of several knowledge-based and corpus-based methods carried out on a large data set of 1,000 deadlines, they tried to identify the methods that work even best for the annotation of emotions since precision of this work is very low.

Zornitsa Kozareva et al[18]focused on headline emotion classification approach based on frequency and co-occurrence information collected from the World Wide Web in 2007. This approach is based on the hypothesis that group of words which co-occur together across many documents with a given emotion are highly probable to express the same emotion. Binary text classification method is used here. Emotion classification is a challenging and difficult task in Natural Language Processing. For the first attempt to detect the amount of angry, fear, sadness, surprise, disgust and joy emotions, they have presented a simple web co-occurrence approach. They have combined the frequency count information of three search engines and measured the Mutual Information score between a bag of content words and emotion. According to the yielded results, the presented approach can determine whether one emotion is predominant or not, and most of the correct sentiment assignments correspond to the negative emotions since they did not consider the impact of valence shifter.

Rada Mihalcea et al[6] introduced the technique of Affective Text Mining in the year 2007. In this paper, the connection between emotions and lexical semantics is used for the classification of emotions and valence (positive/negative polarity). Here each word is first rated separately for each emotion and for valence. Next, the main subject rating was boosted. Contrasts and accentuations between "good" or "bad" were detected, making it possible to identify surprising good or bad news. The system also takes into account: human will (as opposed to illness or natural disasters). Fine Grained (Pearson measure) and Coarse Grained Analysis are used here. The results obtained from this work indicate that the task of emotion annotation is difficult. Although the Pearson correlation for the inter-tagger agreement is not particularly high, the gap between the results obtained by the systems and the upper bound represented by the annotator agreement suggests that there is room for future improvements.

Minqing Hu and Bing Liu[9] introduced the approach of Mining Opinion Features in Customer Reviews in 2004. This paper mainly focused on mining opinion/product features that the reviewers have given judgement. A number of techniques are presented to mine such features. The experimental results show that these approaches are highly effective. Here the system performs summarization in two main steps: feature extraction and opinion orientation identification. The inputs to the system are a product name and an entry page for all the reviews of the product. The output is the summary of reviews. Here, pre-processing includes the deletion of stop words, stemming and fuzzy matching. Symbolic Approach and Statistical Approach are used here. In this paper, they proposed a number of techniques for mining opinion features from product reviews based on data mining and natural language processing methods. But grouping features according to the strength of the opinions that have not been expressed on them. This will further improve the feature extraction and the subsequent summarization.

3. SOCIAL AFFECTIVE TEXT MINING

Consider an online text collection D associated with a dictionary W, and a set of predefined emotions E. In particular, each document d in D consists of a number of words w, and a set of emotion labels. Frequency count of each emotion is also collected by the social websites. The main objective here is to accurately model the connections between words and emotion, and enhance the potential of its related tasks such as emotion prediction. To achieve this, two models are presented here. 1) emotion-term model that uses Naive Bayes to model social emotion and affective terms via their co occurrences and 2) a LDA topic model which utilizes the term co occurrence information within a document and discovers the inherent topics within affective text. Then, later on the proposed emotion-topic model is described that can jointly estimate the hidden document topics and emotion distributions in an integrated probabilistic model.

3.1 Emotion Term Model

Emotion Term Model is mainly used to model the word-emotion association, which follows the Naive Bayes method by assuming words are independently generated from social emotion labels[14]. This model is based on Maximum Likelihood Estimation (MLE) where co occurrence count between words and emotions are estimated for all the documents.

3.2 Topic Model

The topic model used in this work is LDA which is one of the most successful models. LDA overcomes the different problems faced by other models like pLSI by introducing a Dirichlet prior over topics and words. LDA can only discover the topics from document and cannot bridge the connection between social emotion and affective text. In the first study of LDA, Blei et al. [8] proposed a convexity-based variation inference method for inference and parameter estimation under LDA. In this work, an alternative parameter estimation method, Gibbs sampling, is used[14].

3.3 Emotion-Topic Model

Emotion-term model only treats different terms individually and cannot discover the circumstantial information within the document. It is more sensible to associate emotions with a specific topic instead of considering only a single term[9]. While topic model utilizes the contextual information within the documents, it stalls to utilize the emotional distribution to provide the topic generation. Here a new approach called emotion-topic model is used. The emotion-topic model overcomes these drawbacks by introducing an additional emotion generation layer to Latent Dirichlet Allocation. The algorithm works by randomly assigning all the words to emotions and topics[14]. Then repeat Gibbs sampling on each word in the document collection. This sampling process is repeated for N iterations when the stop condition is met.

4. THE PROPOSED SYSTEM

In the proposed system, to increase the efficiency of the processes a joint Emotion topic model for social affective text mining which introduces an additional layer of emotion modeling in to Latent Dirichlet Allocation is used. The proposed emotion topic model allows inferring a number of conditional probabilities of different documents, the probabilities of latent topics given an emotion, and that of terms given a topic. The objective is to accurately model the connections between words and emotions, and improve the potential 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. The four modules used in emotion mining are 1)Admin Process 2)Latent topic generation and processing 3)Extraction and Optimization Process 4)Social Affective Text Mining and Emotion Prediction.

4.1 Admin Process

The administrator collects large amount of URL's and provides it to the server. The server will cross check whether the given URL contents are already present in the database. If the contents are not there, the server will extract and optimize the contents and finally check the vocabulary to find any stop words if present. The emotions in the content will be then categorized and stored in the database. And finally, from the categorized content the emotion will be predicted based on its probability. Emotion-topic model is used here inorder to find the emotion contained in a huge document based on LDA.

4.2 Latent Topic Generation

Here latent topics are generated for each of the identified emotion. As the quantity and quality of the latent topics increases, the efficiency of affective text mining also increases[14]. After collecting and categorizing each latent topics based on different emotions, it is stored in the database. This will be later used to check with the extracted content.

4.3 Extraction And Optimization

Here the title and main body of each article or document is extracted. The access of its main body provides the basis for modeling latent topics and helps alleviate the issue of data sparseness. First we have to segment all the words for each article. Then 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[11]. Finally remove all the stop words that represent no meaning related to any of the specified emotions and thus optimizing the content.

4.4 Social Affective Text Mining And Emotion Prediction

The Emotion Prediction process works as follows. Consider an online text document D, associated with a vocabulary W, and a set of already defined emotions E. Each document d, included in 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 find the frequency count of each word w .Here we are comparing the extracted and optimized content with the already endowed latent topics that relates to each emotion. Based on the result we are predicting which emotion the particular content represents. The user can also provide emotion and based on which the contents will be displayed. Here the connections between words and emotions are established successfully based on topic modelling and hence the performance of its related tasks such as emotion prediction also improved.

5. EVALUATION METRICS

Identifying emotion in text is a type of text classification problem. There are many measures used in text classification. Here, the most commonly used measures are precision, recall, and F-measure. The common metric of interest in evaluating classification systems is how accurate the system is. More specifically, what portion of data belonging to known classes (labelled data) is correctly assigned to those classes. Four datasets were used for the purpose of detecting emotions in text, with sentence-level emotion annotations, used for evaluating emotion identification techniques (news headlines, reports on personal experiences and children's fairy tales). Classification accuracy is usually measured in terms of precision and recall[15]. These two basic measures are explained here for a given document. These are computed as follows:

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

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

True Positive (TP) refers to the number of examples that are classified correctly as belonging to the class, while False Positive (FP) stands for the number of incorrectly classified examples. False Negative (FN) is the number of examples we incorrectly classify as negative [15]. To make it easy to understand, a binary classification problem is taken as an example. There is a positive class and a negative class for a binary classifier. A 2-by-2 confusion matrix shown in Figure 5.1 shows the number of documents predicted correctly and incorrectly into the two classes.

Table 5.1 2-by-2 confusion matrix

		Predicted Emotion	
		Positive	Negative
Actual Emotion	Positive	a	b
	Negative	c	d

The definition of four values is as follows for a given class:
a – number of documents correctly assigned to the class (True Positive)
b – number of documents incorrectly assigned to the class (False Negative)
c – number of documents incorrectly rejected from the class (False Positive)
d – number of documents correctly rejected from the class (True Negative)

Based on these values, (a + b) are the number of documents which truly belong to the positive class. In contrast, (c + d) documents belong to the negative class in the confusion matrix. Precision and recall performance measures are defined and computed from these values.

However, in general precision and recall are not taken into account alone due to its variability. The variability may mislead some of the performance measures. For this reason, a new alternative is used called F-measure[13]. It is used as a metric for effectiveness of classification. The F-measure is defined as follows:

$$\text{F-measure} = \frac{\text{2 x Precision x Recall}}{\text{Precision + Recall}}$$

The F-measure was designed to balance weights of precision and recall. The F-measure values are in the interval (0, 1) and larger F-measure values correspond to higher classification quality. The measure is a popular metric for evaluating classification systems and is most often used to compare the performance of classifiers [13].

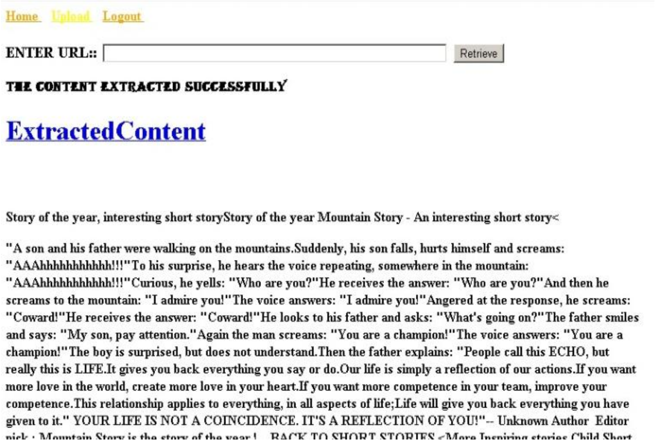
5.1 Text Similarity Measures

A number of different methods have been devised and compared for evaluating the semantic similarity of documents. Documents can be represented as vectors through VSM (Vector Space Model) [16]. The vector-based semantic representation of documents leads to a lot of benefits to take advantage of various linear algebra operations. On top of that, terms, sentences, paragraphs, and documents can be represented in a uniform way on the vector space [5]. Similarity between sentences is computed in one of two ways. The first technique is the dot-product calculating the Euclidean distance between the two vectors and the second method is the cosine similarity taking a measure of the angle between the vectors [16]. The latter technique emphasizes the directions of vectors rather than the lengths of them. Euclidean distance is a generally used measure of similarity in the area of image data, while cosine similarity is widely exploited in the realm of text data [2]. In particular, cosine similarity is commonly used in some supervised learning algorithms for document categorization [6].

The similarity in a lower dimensional semantic space is measured by the standard cosine value. More importantly, a vector space model allows words, synsets, and sentences to be compared with each other as well as to be represented homogeneously. In practice, the cosine of the angle between an input vector (input sentence) and an emotional vector (emotional synsets) is computed to identify which emotion the sentence connotes [3]. Linear combinations of emotional synonym vectors are used in order to form a vector representation of one emotion. The more closely two vectors are related semantically, the higher their cosine value is. As an example, assume that the input sentence is “In a cottage in a large forest, I was alone for a while”. If the calculated cosine values between the input and “joy” vector, and “fear” vector are 0.3 and 0.7, respectively, it is concluded that the input sentence implies the fear emotion. if the cosine similarity does not exceed a threshold, the input sentence is labelled as “neutral”, the absence of emotion. Otherwise, it is labelled with one emotion associated with the closest emotional vector having the highest similarity value.

6. RESULT AND ANALYSIS

In order to predict the emotion contained within a document, LDA under topic modelling is used. Here, the latent topics are identified from a given input document or article and then different classification algorithms are used to convert pseudo-documents into predefined categories. Then dimension reduction is done which improves the accuracy of emotion prediction process. This is followed by extraction and optimization process which removes all the stop words that represent no meaning related to any of the specified emotions and thus optimizing the content. The extraction and optimization process is shown in figure 6.1 and 6.2 respectively. Admin have to login in order to perform extraction and optimization. The administrator collects large amount of URL’s and provides it to the server. The server will cross check whether the given URL contents are already present in the database. If the contents are not there, the server will extract and optimize the contents and finally check the vocabulary to find any stop words if present. The emotions in the content will be then categorized and stored in the database. And finally, from the categorized content the emotion will be predicted based on its probability. Text based emotion detection systems have gone a step beyond simple word matching by performing a semantic analysis of the text. Techniques for detecting emotions in text have been applied to different application domains and the result shows that connection between emotion and affective terms are established successfully. The main goal of this affective text mining is to find the closest emotion category in a given document and also to predict a single emotional label given an input sentence.



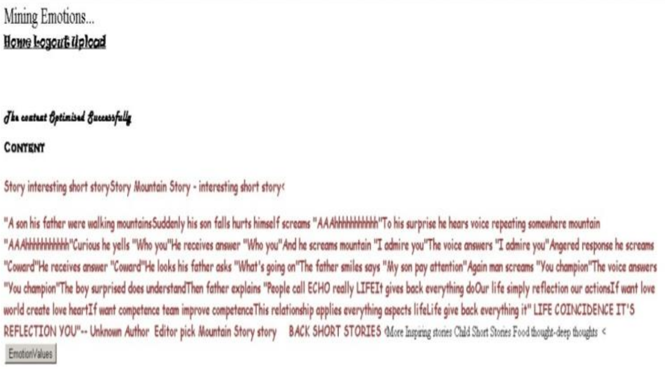
THE CONTENT EXTRACTED SUCCESSFULLY

ExtractedContent

Story of the year, interesting short storyStory of the year Mountain Story - An interesting short story<

"A son and his father were walking on the mountains.Suddenly, his son falls, hurts himself and screams: "AAAhhhhhhhhhhh!!!"To his surprise, he hears the voice repeating, somewhere in the mountain: "AAAhhhhhhhhhhh!!!"Curious, he yells: "Who are you?"He receives the answer: "Who are you?"And then he screams to the mountain: "I admire you!"The voice answers: "I admire you!"Angered at the response, he screams: "Coward!"He receives the answer: "Coward!"He looks to his father and asks: "What's going on?"The father smiles and says: "My son, pay attention."Again the man screams: "You are a champion!"The voice answers: "You are a champion!"The boy is surprised, but does not understand.Then the father explains: "People call this ECHO, but really this is LIFE.It gives you back everything you say or do.Our life is simply a reflection of our actions.If you want more love in the world, create more love in your heart.If you want more competence in your team, improve your competence.This relationship applies to everything, in all aspects of life;Life will give you back everything you have given to it." YOUR LIFE IS NOT A COINCIDENCE. IT'S A REFLECTION OF YOU!"-- Unknown Author Editor pick : Mountain Story is the story of the year ! BACK TO SHORT STORIES <More Inspiring stories Child Short

Figure 6.1 Content Extraction Process



The content Optimised Successfully

CONTENT

Story interesting short storyStory Mountain Story - interesting short story<

"A son his father were walking mountainsSuddenly his son falls hurts himself screams "AAAhhhhhhhhhh"To his surprise he hears voice repeating somewhere mountain "AAAhhhhhhhhhh"Curious he yells "Who you"He receives answer "Who you"And he screams mountain "I admire you"The voice answers "I admire you"Angered response he screams "Coward"He receives answer "Coward"He looks his father asks "What's going on"The father smiles says "My son pay attention"Again man screams "You champion"The voice answers "You champion"The boy surprised does understandThen father explains "People call ECHO really LIFEIt gives back everything doOur life simply reflection our actionsIf want love world create love heartIf want competence team improve competenceThis relationship applies everything aspects lifeLife give back everything it" LIFE COINCIDENCE IT'S REFLECTION YOU"-- Unknown Author Editor pick Mountain Story story BACK SHORT STORIES <More Inspiring stories Child Short Stories Food thought-deep thoughts <

Figure 6.2 Content Optimization Process

7. CONCLUSION AND FUTURE WORK

In this paper, an empirical study of the text-based emotion prediction is done. From this study, a new problem called social affective text mining is analyzed, which aims to discover and model the connections between online documents and user-generated social emotions. To the end, a new joint emotion-topic model by augmenting Latent Dirichlet Allocation (LDA) with an intermediate layer for emotion modeling is proposed. Rather than emotion-term model that treats each term in the document individually and LDA topic model that utilizes only the text co occurrence information, emotion-topic model overcomes these drawbacks by introducing an additional emotion generation layer to Latent Dirichlet Allocation. The algorithm works by randomly assigning all the words to emotions and topics. Then repeat Gibbs sampling on each word in the document collection. This sampling process is repeated for N iterations when the stop condition is met. Experimental result shows that the model is effective in extracting the meaningful latent topics, and also significantly improves the performance of social emotion prediction compared with the baseline emotion-term model and multiclass SVM. As for future work, we are planning to evaluate our model with a larger scale of online document collections, and apply the model to other applications such as emotion-aware recommendation of advertisements, songs,
and so on.

REFERENCES

1. A.-M. Popescu and O. Etzioni, “Extracting Product Features and Opinions from Reviews,” Proc. Joint Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP ’05), pp. 339-346, 2005.
2. Bingham E. and Mannila H., “Random projection in dimensionality reduction”, Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 26-29 August 2001.
3. C. Strapparava and R. Mihalcea, “Learning to Identify Emotions in Text,” Proc. 23rd Ann. ACM Symp. Applied Computing (SAC ’08), pp. 1556-1560, 2008.
4. 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.

5. Carlo Strapparava and Alessandro Valitutti. "WordNet-Affect: an Affective Extension of WordNet", in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pp. 1083-1086, Lisbon, May 2004.
6. Carlo Strapparava and R. Mihalcea, “Semeval-2007 Task 14: Affective Text,” Proc. Fourth Int’l Workshop Semantic Evaluations (SemEval ’07), pp. 70-74, 2007.
7. Chung-Hsien Wu and Xu, S. Bao, “Emotion recognition from text using semantic labels and separable mixture models,” Proc. Fifth Int’l Conf. Language Resources and Evaluation (LREC ’05), 2005.
8. G. Mishne, K. Balog, M. de Rijke, and B. Ernsting, “Moodviews: Tracking and Searching Mood-Annotated Blog Posts,” Proc. Int’l AAAI Conf. Weblogs and Social Media (ICWSM ’07), 2007.
9. Minqing 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.
10. Onana Frunza, Diana Inkpen, And Thomas Tran, “A Machine-learning Approach for Identifying Disease Treatment Relations In Short Text”, IEEE Transactions on Knowledge and Data Engineering, Vol 23, No. 6, 2011.
11. R. Tokuhisa, K. Inui, and Y. Matsumoto, “Emotion Classification Using Massive Examples Extracted From The Web,” Proc. 22nd Int’l Conf. Computational Linguistics (Coling ’08), pp. 881-888, 2008.
12. S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fuku-shinna, “Mining Product Reputations on the Web,” Proc. Eighth ACM SIGKDDInt’l Conf. Knowledge Discovery and Data Mining (SIGKDD’02), pp. 341-349, 2002.
13. S.M. Kim, A. Valitutti, and R.A. Calvo. Evaluation of Unsupervised Emotion Models to Textual Affect Recognition. Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, pp. 62-70, Los Angeles, California, June 2010.
14. 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.
15. Sophia Yat Mei Lee, Ying Chen and Chu-Ren Huang, “A Text-driven Rule-based System for Emotion Detection,” Proc. Joint Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP ’05), pp. 579-586, 2005.
16. Sunghwan Mac Kim, “Recognising Emotions and Sentiments in Text,” The University of Sydney, April 2011.
17. 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.
18. Zornitsa Kozareva, Borja Navarro, Sonia V´azquez, Andr´es Montoyo, “A Headline Emotion Classification through Web Information,” Proc. IEEE/WIC/ACM Int’l Conf. Web Intelligence, pp. 275-278, 2007.