Document Level Emotion Tagging: Machine Learning and Resource Based Approach

Dipankar Das and Sivaji Bandyopadhyay

Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
dipankar.dipnil2005@gmail.com, sivaji_cse_ju@yahoo.com

Abstract. The present task involves the identification of emotions from Bengali blog documents using two separate approaches. The first one is a machine learning approach that accumulates document level information from sentences obtained from word level granular detail whereas the second one is a resource based approach that considers the Bengali WordNet Affect, the word level Bengali affective lexical resource. In the first approach, the Support Vector Machine (SVM) classifier is employed to perform the word level classification. Sense weight based average scoring technique determines the sentential emotion scores based on the word level emotion tagged constituents. The cumulative summation of sentential emotion scores is assigned to each document considering the combinations of various heuristic features. The second one implements a majority based approach to classify a given document considering the Bengali WordNet Affect lists. Instead of assigning a single emotion tag to a document, in both approaches, the best two emotion tags are assigned to each document according to the ordered emotion scores obtained. By applying the best feature combination acquired from the development set, the evaluation of 110 test documents yields the average *F-Scores* of 59.50% and 51.07% for the two approaches respectively with respect to all emotion classes.

Keywords. Natural language processing, computational linguistics, text, blog, document, WordNet Affect, sense weight score, CRF, SVM, emotion tagging, heuristic features.

Etiquetación de emociones a nivel de documento: aprendizaje automático y un método basado en recursos

Resumen. El objetivo de este trabajo es identificar las emociones en documentos escritos en bengalí extraídos de un blog usando dos enfoques distintos. El primer enfoque es aprendizaje automático en el cual se acumula la información de los documentos a partir de las oraciones obtenidas a través de análisis de palabras, es decir, en el nivel más granular, mientras que el segundo enfoque está basado en recursos de los cuales usamos el Bengalí WordNet Affect —un recurso léxico que incluye palabras del bengalí etiquetadas con emociones. En el primer enfoque, la máquina de soporte vectorial (Support Vector Machine, SVM) se usa para la clasificación a nivel de palabras. El valor afectivo de las oraciones se calcula según la técnica basada en promediar los puntajes de pesos asignados a los significados de palabras etiquetadas con emociones en estas oraciones. La suma acumulada de los puntajes afectivos de las oraciones se asigna a cada documento tomando en cuenta diversas características heurísticas. El segundo enfoque implementa el método basado en mayoría para clasificar un documento dado considerando las listas del Bengalí WordNet Affect. En ambos enfoques, en vez de asignar una única etiqueta afectiva a un documento dado, las dos mejores etiquetas afectivas se asignan a cada documento según los puntajes afectivos obtenidos ordenados. Usando la combinación de las mejores características obtenida del conjunto de desarrollo, al evaluar 110 documentos de prueba resulta un valor promedio de la métrica *F-score* en los dos enfoques 59,50% y 51,07% respectivamente para toda clase de emociones.

Palabras clave. Procesamiento de lenguaje natural, lingüística computacional, texto, blog, documento, WordNet Affect, puntaje de peso de significado, campo aleatorio condicional (Conditional Random Field, CRF), máquina de soporte vectorial (Support Vector Machine, SVM), etiquetas afectivas, características heurísticas.

1 Introduction

A text contains not only informative contents, but also more or less attitudinal private information including emotional states. In psychology and common use, emotion is an aspect of a person's mental state of being, normally based in or tied to the person's internal (physical) and external (social) sensory feeling [41]. Though the human emotion described in texts is an important cue for our daily communication, the identification of emotional state from texts is not an easy task as emotion is not open to any objective observation or verification [31]. Moreover, the same textual content can be presented with different emotional slants [18]. Hence, the identification of the emotional state from text is really a challenging issue.

On the other hand, emotion analysis is a recent sub-discipline emerged at the crossroads of information retrieval [33] and computational linguistics [38]. Information is concerned not only with the topic of a document, but also with the emotion it expresses. The classification of reviews [36] or newspaper articles [24], Question Answering systems [2] and modern Information Retrieval systems [30] are increasingly incorporating emotion analysis within their scope. It is sometimes important to track users' emotion expressed in online forums, blogs or twitters for different applications such as sentiment review, customer management, stock exchange prediction, etc. Blogs are one of the important communicative and informative repositories of text-based emotional contents in the Web 2.0 [24]. Researches on emotion show that blogs play the role of a substrate to analyze the reactions of different emotional enzymes. Especially, the blog posts contain instant views, updated views or influenced views regarding single or multiple topics. Many blogs act as an online diary of the bloggers for reporting the blogger's daily activities and surroundings. Sometimes, the blog posts are annotated by other bloggers.

Several efforts have been made by the Natural Language Processing (NLP) researchers to identify emotions at different levels of granularity such as the word, sentence or document level [9] [12, 21]. It is said that sentiment or emotion is typically a localized phenomenon that is more appropriately computed at the paragraph, sentence or entity level [25]. In general, the sentence level emotion identification plays an important role to track emotions or to find out the cues for generating such emotions or to identify them properly. Sentences as the basic information units of any document identify the overall document level emotion whereas emotions of individual sentences in the documents are based on the emotions expressed by the word(s) [4]. In the present task, each of the blog documents is annotated with Ekman's [16] six basic emotion tags based on the heuristically produced knowledge of the sentence level emotion tags [13]. On the other hand, the assignment of sentential emotion tags is carried out based on the word level emotion tagged constituents [8].

It is well known that a large collection of blog data is suitable for any machine learning framework. Earlier attempts were carried out using different unsupervised, supervised and semi-supervised strategies for identifying and classifying emotions [32]. A supervised classifier based on *Support Vector Machine* (SVM) [7] was used on the blog data to classify the documents according to the mood of the author during writing [28, 29]. The authors used emoticons in LiveJournal posts to train a mood classifier at the document level. In another experiment [39], the researchers have performed the emotion classification task on web blog corpora using SVM and Conditional Random Field (CRF) [22] based machine learning techniques. It was observed that the CRF-based classifier outperforms the SVM classifier in the case of the document level emotion detection. Instead of using a CRF-based classifier, the present task incorporates Support Vector Machine (SVM) for word level emotion tagging. It is observed that SVM outperforms CRF in word level emotion tagging. The reason may be that the CRF used for a sequence labeling problem suffers in the tagging of emotions due to discrete word tokens that are well considered by SVM.

The present work involves a weight based scoring technique to identify the document level emotion tags. At first, the sense weight based average scoring technique [8] is applied for assigning sentential emotion tags based on the word level emotion tagged constituents. Simple rule based techniques are employed for handling the negations present at the sentence level. Finally, the document level emotion tagging is carried out based on the emotion scores obtained for the sentences along with some heuristic feature combinations (e.g. the emotion tag of the title sentence or the end sentence of the topic section, emotion tags assigned to the overall topic, most frequent emotion tags expressed in the user comment portions of a document, identical emotions that appear in the longest series of tagged sentences etc.). The best two emotion tags are assigned to a document based on the ordered maximum emotion scores obtained. The development set gives the best average *F-Score* of 59.32% after applying the possible feature combinations. Evaluation is carried out against the best two annotated emotion tags of 110 test documents containing 1298 comment sections of the bloggers. The average *F-Score* of 59.50% has been achieved with respect to all emotion classes.

The classification of news articles according to the readers' emotions instead of the authors' ones is presented in [24]. In the present task, the perspective of reader's emotion is considered only for judging over the emotional counterpart. Yahoo! Kimo Blog corpora were used to build emotion lexicons [39, 40]. In the latter studies, emoticons were used to identify emotions associated with textual keywords. The emoticons of the Bengali blog documents are also considered in the present task.

Several lexical resources have been developed for English in the past to help opinion, sentiment or emotion analysis [1, 5, 17, 35]. Major studies on opinion mining and sentiment analysis have been attempted with more focused perspectives rather than fine-grained emotions. On the other hand, a rapidly growing number of web users from multilingual communities focus their attention on improving the multilingual search engines on the basis of sentiment or emotion. The analysis of emotion or sentiment requires some basic resource. An affective lexicon is one of the primary resources to start with as the identification of direct emotion words in the lexicon lookup approach. A lexicon that provides appraisal attributes for terms was constructed and the features were used for emotion classification [37]. The features along with the bag-of-words model give 90.2% accuracy. UPAR7 [3], a rule-based system uses a combination of WordNet Affect [35] and SentiWordNet [17], which were semi-automatically enriched with the original trial data provided during the Semeval task [34]. SWAT [19] is another supervised system that uses a unigram model trained to annotate emotional content. Synonym expansion on the emotion label words has used the Roget Thesaurus. But the effectiveness of a lexical resource for emotion analysis in Bengali motivates us to find the usage of such resource by creating a majority based classifier for emotion prediction. Though Bengali is a very resource-constrained language, a recently developed Bengali *WordNet Affect Lists* (*BengWAL*) [10] have been used in the present task to determine the directly stated emotion words. Apart from the stop word removal, we have applied stemming for each word in the document and looked up for the word in the *BengWAL*.

We have assigned emotions to the documents based on the majority presence of words with those emotions in the documents. The best two emotion tags are assigned to a document based on the emotion classes for which the maximum and the next to maximum number of emotion words appear in that document. Preprocessing steps like stemming and lemmatization have been found to be detrimental to classification accuracy [23].

In the present approach, stemming plays a contributory role in emotion classification to cope with the morphological enrichment properties of Bengali. Experiments have been conducted with *SentiWordNet* [15] to study whether such a resource can aid in the sentiment classification or not. In general, these studies observe that resource-based sentiment classifiers are less effective compared to machine learning-based approaches. It has also been observed that the machine learning approach (the average *F-Score* of 59.50%) outperforms the lexicon-based approach (the average *F-Score* of 51.07%) with respect to all emotion classes. Nevertheless, we believe that the lexical resources created for such resource-based systems can be utilized for developing the decisive feature vectors.

2 Corpus Preparation

The mode of language technology has changed dramatically since the last few years with the web being used as a data source in a wide range of research activities. India is a multilingual country with a diverse cultural heritage. Indian languages are resource-constrained. Hence, an emotion annotated Bengali blog corpus has been developed manually [11] to accomplish our present goal. Each sentence of the corpus is annotated with emotional components such as emotional expression (word/phrase), intensity, associated holder and topic(s). Any of the Ekman's six basic universal emotion classes (*anger, disgust, fear, happy*, *sad* and *surprise*) along with three types of intensities (*high*, *general* and *low*) are considered for the sentence level annotation. To the best of our knowledge, Ekman's six basic emotions are considered as the universal emotions, since all other non-basic emotions can be defined in terms of the six basic categories [16]. A sentence level emotion tagging system has been implemented and evaluated on this annotated corpus. The expression level as well as the sentence level emotion annotation details are mentioned in [11]. Instead of using the emotional information regarding expression or intensities, we have considered only the sentential emotion tags in our present task. But, this annotated corpus does not contain any information regarding document level emotion tagging. Hence, we have collected the blog documents from the web blog archive. The format of the blog document is shown in Figure 1. Each of the blog documents is assigned a unique identifier (*docid#*) followed by a section devoted to the topic along with several sections devoted to various users' comments. Each comment section consists of several nested and overlapped sub-sections that also contain the bloggers' comments. But, each of the comment sections of an individual blogger is uniquely identified by the notion of the section identification number (*secid#*). Each section contains information regarding identification number of the blog user (*uid#*) and the associated timestamp (*tid#*). Recently, another study of emotional changes has been conducted on this corpus for tracking emotions of the bloggers [14].

The Bengali blog documents contain a significant number of emoticons that play a contributory role in the document level emotion tagging. A total of 24,678 words are used in the present task. We have considered a total of 205 blog documents in which 95 and 110 documents are treated as the development and test sets respectively. We have used a sentence level emotion tagger for assigning emotions to the sentences of these blog documents. The distribution of documents, sentence and words in the training, development and test corpora is given in Table 1a.

```
-<DOC docid = xyz>
    +<Topic>.... </Topic>
    -<User Comments id=UC1>
        -<U uid=1, tid=t1, secid=UC1>....
            -<U uid=2, tid=t2, secid=UC1.1>...</U>
            -<U uid=3, tid=t3, secid=UC1.2>...</U>
                -<U uid=1, tid=t4, secid=UC1.2.1>...</U>
            ....
        </U>
    </User Comments>
    +<User Comments id=UC2>
    +<User Comments id=UC3>
    ...
</DOC>
```

Fig. 1. General structure of a Bengali blog document

The tagging of each document with Ekman's six universal emotions was carried out manually. No prior training was provided to the annotators to avoid the bias of personal emotions of the annotators during the emotion annotation process. The annotators were instructed to annotate each sentence of the blog corpus based on some illustrated samples of the annotated sentences. Three annotators identified as A1, A2 and A3 carried out the annotation. The annotation agreement was measured using the standard Cohen's *kappa* coefficient (κ) [6]. It is a statistical measure of inter-rater agreement for qualitative (categorical) items. It measures the agreement between two raters who separately classify items into some mutually exclusive categories.

Each of the blog documents may be annotated with more than one emotion tags. Out of total 205 blog documents, only 78 documents contain a single emotion tag (*Sing*). It has been observed that the emotion pairs such as "*sad-anger*" and "*anger-disgust*" often cause trouble in distinguishing the emotion at the document level. The mixed emotion category (*Mult*) gives poor agreement results as expected. The inter-annotator agreement for the document level emotion tagging is shown in Table 1b whereas the sentence level agreement results can be found in [11].

Table 1a. Distribution of documents (D), sentences (S) and words (W) in training, development and test corpora

Distribution	Sentence level tagger	Document level tagger
Training	700 (S), 9302 (W)	Use of a sentence level tagger
Development	300 (S), 4115 (W)	95 (D), 1067 (S), 11,534 (W)
Test	200 (S), 3087 (W)	110 (D), 1304 (S), 13,144 (W)

Table 1b. Inter-annotator agreement using *kappa* (κ)

Pair of annotators	Agreement, kappa
A1↔A2	0.85 (*Sing*), 0.43 (*Mult*)
A1↔A3	0.92 (*Sing*), 0.48 (*Mult*)
A2↔A3	0.95 (*Sing*), 0.57 (*Mult*)

3 Word to Sentence Level Emotion Tagging

A small portion of the blog corpus [11] containing 1,200 sentences and 16,504 word tokens is considered for word level emotion tagging in the SVM-based machine learning framework. A set of standard preprocessing techniques is carried out, *viz.*, *tokenizing, stemming* and *stop word removal* using the tools developed in the laboratory. The word level, as well as the sentence level emotion-annotated information helps in measuring the performance of the system at both levels of granularity. Out of the total 1,200 sentences collected from 14 different blog documents, 700 sentences are considered for training. The development set and the test set consist of 300 and 200 sentences respectively. The results reported for word level emotion tagging are based on five-fold cross validation.

3.1 Word Level Classification

In the present task, the technique adopted for the word level emotion classification is similar to the approach reported in [9]. However, instead of using Conditional Random Field (CRF), the Support Vector Machine (SVM) based classifier is used in the present task to classify each word into any of the Ekman's six emotion categories. The training and classification processes for SVM have been carried out by YamCha toolkit[1] and TinySVM-0.07 [2] respectively.

Features play a crucial rule in any machine-learning framework. Therefore, among 10 active singleton features of [9] [12], 9 features have been employed to accomplish the current task. The *Bengali WordNet Affect* [10] has been used in this task instead of Bengali *SentiWordNet* to identify the features for the emotion words. Different unigram and bi-gram context features (word level as well as POS tag level) have been applied. The features are as follows:

– POS information (*adjective*, *verb*, *noun*, *adverb*),
– Words of the title sentence or the first sentence in the topic section and bloggers' comments,
– Emotion words of *Bengali WordNet Affect* (e.g., সুন্দর *sundar* [*beautiful*]),
– Reduplication (e.g., *bhallo bhallo* [*good good*], *khokhono khokhono* [*when when*] etc.),
– Question words (*ki* [*what*], *keno* [*why*] etc.),
– Colloquial/Foreign words (e.g. ক্ষ্যামা *kshyama* [*pardon*]) and foreign words (e.g., *Thanks, gossya* [*anger*] etc.),
– Special punctuation symbols (!, @, ?..),
– Quoted sentence ("*tumi khub bhalo lok*" [*you are 2 good man*]),
– Sentence length (>=8, <15),
– Emoticons (☺, 😐, ☹ ..).

A CRF-based classifier is generally applied to accomplish the classification task for a sequence labeling problem. It fulfills a word level

[1] http://chasen-org/~taku/software/yamcha/
[2] http://chasen.org/~taku/software/TinySVM/

classification task with a significant loss of word level emotional constituents. As SVM gives better performance in a discrete (e.g. word) information tagging, the improvement is observed in the SVM-based word level emotion classification task. The comparative results of the earlier system [8] using CRF and the present system using SVM for word level emotion tagging are shown in Table 2. The improvement of the word level emotion tagging system is reflected in the sentential emotion tagging.

3.2 Sentential Emotion Tagging

The default emotion tag weights [9] [12] for six emotion types are considered. The six basic words "*happy*", "*sad*", "*anger*", "*disgust*", "*fear*" and "*surprise*" are chosen as the *seed words* corresponding to each emotion type. The *positive* and *negative* scores in the English *SentiWordNet* [17] for each synset in which each of these *seed words* appear are retrieved and the average of the scores is fixed as the *Sense_Tag_Weight* (STW) of that particular emotion type.

Table 2. Word level accuracies (in %) of CRF and SVM

Emotion Class (#Words)	Test Set	
	CRF	SVM
Happy(106)	67.67	69.55
Sad(143)	63.12	65.34
Anger(70)	51.00	56.15
Disgust(65)	49.75	53.35
Fear(37)	52.46	54.78
Surprise(204)	68.23	69.37

The present work differs from the approach in [9] in the sense of assigning emotion scores and sentential emotion tags to the blog sentences. In the present method, the emotion tagged words have been considered instead of depending on the fixed emotion tag weights assigned to the words of a specific emotion class. For supporting the task, a Bengali *SentiWordNet* is being developed by replacing each word entry in the synonymous set of the English *SentiWordNet* by its possible set of Bengali synsets using a synset-based English to Bengali bilingual dictionary being developed as a part of the EILMT Project[3].

Each word tagged with a particular emotion type is searched in the Bengali *SentiWordNet* and the *positive* and *negative* scores of the word are retrieved from the *SentiWordNet*. The average of the scores is fixed as the *Sense_Tag_Weight* (STW) for the emotion tag assigned to that word. If an emotion tagged word is not found in the Bengali *SentiWordNet*, the default weight calculated earlier is assigned to that word. The total *Sense_Tag_Weight* (STW) for each emotion tag *i* is calculated by summing up the STWs of all assigned emotion tags with type *i*. Stemming is included during the searching process. Bengali, like any other Indian languages, is very rich morphologically. Different suffixes may be attached to a word depending on various features (e.g. the features for a Bengali verb are Tense, Aspect, and Person). An in-house Bengali stemmer uses a suffix list to identify the stem form of the word.

Table 3. Sentence level accuracies (in %)

Emotion Class (#Sentences)	Test Set	
	CRF	SVM
Happy(40)	65.28	66.05
Sad(41)	66.42	68.12
Anger(32)	60.28	62.77
Disgust(21)	52.18	53.54
Fear(23)	57.14	60.11
Surprise(43)	66.45	69.82

Apart from the search technique, the sentential emotion tagging is carried out in the manner reported in [12]. Each sentence is assigned a *Sense_Weight_Score* (SWS) for each emotion type. The weight is calculated by dividing the total STW of all occurrences of that emotion tag in the sentence by the total STW of all types of emotion tags present in that sentence. The sentence is assigned the emotion tags for which the sentence level *Sense_Weight_Score* (SWS) is highest. The sentences are tagged as *neutral* type, if for all

[3] English to Indian Languages Machine Translation (EILMT) is a TDIL project undertaken by the consortium of different premier institutes and sponsored by MCIT, Govt. of India.

emotion tags, the total of *Sense_Weight_Scores* (SWS) produces zero (0) emotion score. The post-processing strategies [12] related to negative words have been incorporated in the present system. The comparative results of the CRF- and SVM-based models for sentence level emotion tagging is shown in Table 3.

The presence of negative words and the number of their occurrences are significant in assigning the final emotion tag to a sentence. The sentences have been passed through the post-processing system for handling the negative words [12]. The application of a rule-based negation handling technique improves the performance of the system by 2.05% in terms of *F-Score* on the development set. The rules are finally applied on the test set.

4 Sentence to Document Level Emotion Tagging

Assigning a single emotion tag to a particular document does not always bear the actual emotions present in that document. This module identifies six document level emotion tags and their associated weights based on the sentence level emotion tags along with the contribution of heuristic features.

Emotion_Weight_Score (EWS) based technique applied on the sentence level emotion tags produces six possible emotion scores for a document. However, the evaluation is carried out for the best two emotion tags only [13].

4.1 Calculation of Document Level Emotion Tag Weights

Each document is assigned an *Emotion_Weight_Score* (EWS) for each of the six emotion tags. In general, the document level *Emotion_Weight_Score* (EWS) for a particular emotion tag is calculated by summing up the total *Sense_Weight_Scores* (SWS) of all occurrences of the sentential tags corresponding to that emotion category, i.e., $EWS_i = \sum SWS_i$, where SWS_i is the sentence level *Sense_Weight_Score* (SWS) for the emotion tag *i* in the document.

Each document is assigned two document emotion tags DET_i and DET_j, for which the *Emotion_Weight_Scores*, EWS_i is the maximum and EWS_j is the next to maximum, i.e., $DET_i = [\text{Max}_{i=1 \text{ to } 6}(EWS_i)]$ and $DET_j = [\text{Max}_{j=1 \text{ to } 6 \;\&\&\; j \neq i}(EWS_j)]$.

4.2 Evaluation

The emotion tags corresponding to the maximum and the next to maximum *Emotion_Weight_Scores* (EWS) of a document are considered as the probable candidate emotion tags. The set, namely, **GSDT** (Gold Standard Document Tag) contains at most two emotion tags that are assigned to a document in the gold standard annotated corpus and is defined as {*dmax1, dmax2*}. On the other hand, the document level emotion tagging module generates the set **SGDT** (System Generated Document Tag) that contains two probable candidate emotion tags for a document based on their ordered *Emotion_Weight_Scores* (EWS) and the set is defined as {*dmax1´, dmax2´*}. The *F-Score* for each emotion tag pair is measured by considering the number of system generated document tags that match correctly with the annotated tags. The final average *F-Score* is calculated for each emotion class considering any four combinations of the two sets. It has to be mentioned that the performance of the system in terms of *F-Score* has not improved significantly by adding more than two tags in GSDT and SGDT sets in the extended evaluative experiments.

The tagged documents are evaluated against the manually annotated gold standard documents. It is observed that 59.32% average *F-score* has been achieved with these four combinations on a development set of 95 documents with respect to six emotion classes. But, the final document level emotion tagging considers the heuristic features and their different combinations. The corresponding feature combinations that give the best average *F-Score* on the development set is applied to 110 test documents and finally the average *F-Score* of 59.50% has been achieved.

4.3 Experiments over Heuristic Features

Document level emotion identification depends not only on the emotion expressed in the sentential constituents but also on combinations of various characteristic features of that document. Irrespective of the linguistic attributes, a blog itself contains some special inherent features that help in identifying emotions at the sentence level as well as at the document level. In the present task, the following seven active features are identified heuristically and employed in a SVM-based word level classifier. The numeric figures in brackets denote the number of times the corresponding feature has appeared in the development and test sets respectively. The features are defined as follows:

1. Emotion tags of the title sentence (95, 110).
2. Emotion tags of the start and end sentence of the topic sections (95, 110).
3. Emotion tags assigned to the overall topic section (95, 110).
4. Emotion tags that are present in the bloggers' comment sections (1156, 1298).
5. Most frequent emotion tags identified from the document (95, 110).
6. Identical emotions that appear in the longest series of tagged sentences (67, 61) [39].
7. Emotion tags of the largest section among all the user comments' sections (1274, 1322).

Each of the documents contains a title and a topic and hence the frequencies of the first three features are the same as the number of documents. The development and test documents contain a total number of 1,156 and 1,298 bloggers' comment sections.

The emotions reflected inside the comment sections are helpful in predicting the overall emotions at the document level. An emphasis is also given to the frequency of the emotion tags identified at the document level. It is observed that the comment sections are organized in a nested fashion. Hence, this feature enhances the performance of predicting the overall emotions expressed in the documents. The contributions of the features alone and in combination with other features have been evaluated on 95 documents of the development set. The frequencies of different features are shown in Figure 2. It has been observed that the emotion classes *fear* and *disgust* contain less frequent information regarding features (1), (2) and (5). The average *F-score* value is calculated on the development set for each of the features, and moreover, the individual contribution of each feature has been measured. It is found that the contribution of each feature is not uniform and can be fairly distinguished according to the level of importance. For example, the combination of topic *(3)* as well as bloggers' comments *(4)* is identified to be a contributory feature pair as denoted by the experiment id *ii (8)*. Though the contributions of all features and their combinations are not mentioned, some of the important experimental results are shown in Table 4. It is also observed that the emotions expressed in the title of the documents do not always convey the actual emotions expressed inside the documents. An important observation is that, as the number of feature instances varies in the emotion classes, they also have an impact on the document level emotion tagging.

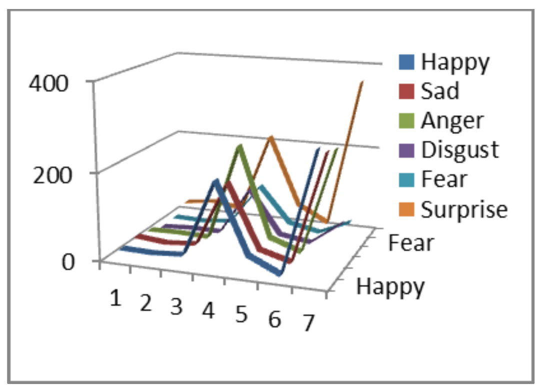

Fig. 2. Frequencies of seven features per emotion class in the development set

5 Resource Based Emotion Tagging

A naive approach for classifying an emotional document is to use the affective terms present in it. In order to find the emotional or affective connotations, a lexical resource is required. In this section, we explain the methodology used for developing and employing such a resource for document level emotion tagging. One of the affective resources is Bengali *WordNet Affect Lists* (*BengWAL*). *BengWAL* is created by exploiting four lexical resources of English namely *WordNet Affect* [35], *SentiWordNet* [17], *VerbNet* [20] and an English-Bengali bilingual dictionary.

5.1 WordNet Affect

The English *WordNet Affect* [35], based on Ekman's six emotion types, is a lexical resource containing information about the emotional words. The *WordNet Affect* is a small lexical resource compared to the complete *WordNet* [27], but its affective annotation helps in emotion analysis. The collection of *WordNet Affect* synsets was provided as a resource for the shared task of *Affective Text* in *SemEval-2007* [34]. A part of the original *WordNet Affect* was fine-grained and re-annotated using six emotional labels like *joy, fear, anger, sadness, disgust, surprise* [16].

It is developed based on *WordNet* domains [26] where each synset is annotated with at least one *domain label*, selected from a set of two hundred labels that are arranged hierarchically. In addition to that, *WordNet Affect* contains an additional hierarchy of the *affective domain labels*. Without considering the problems of the lexical affect representation or the differences between emotions, cognitive states and affects, we have developed the *WordNet Affect* lists in Bengali from the available lists of English *WordNet Affect* [35].

5.2 Development of BengWAL

The development of Bengali *WordNet Affect Lists* (*BengWAL*) consisted of four phases, such as updating, translation, sense disambiguation and evaluation.

Each of the English *WordNet Affect* lists was updated with the equivalent synsets retrieved from the English *SentiWordNet* [17] to reach an adequate number of emotion word entries. The part-of-speech (POS) information for each of the synsets was kept unchanged. Member verbs present in the same *VerbNet* [20] class share common syntactic frames so they were believed to have the same syntactic behavior. Hence, the member verbs belonging to each of the *VerbNet* classes produced an individual verb synset. Each verb present in the six affect lists was updated with the retrieved *VerbNet* synsets if any word level match occurs. We employed a duplicate removal technique that accumulated more emotion bearing words with reduced error from the updated synsets.

Table 4. F-Scores (in %) of heuristic features

Expt.ID	Features and Combinations	Average F-score (in %)
I	(1) Emotion tags of the title sentence	31.12
	(2) Emotion tags of the end sentence of a topic	28.25
	(3) Emotion tags assigned to an overall topic	48.87
	(4) Emotion tags for user comment portions of a document	52.66
	(5) Most frequent emotion tags identified from the document	53.95
	(6) Identical emotions that appear in the longest series of tagged sentences	37.29
	(7) Emotion tags of the largest section among all of the user comments' sections.	35.11
II	(8). i(3)+i(4)	57.32
	(9). i(3)+i(5)	56.55
	(10). i(3)+i(7)	55.42
	(11). i(4)+i(5)	54.87
	(12). i(4)+i(6)	53.25
	(13). i(4)+i(7)	55.57
III	(14). ii(8)+i(6)	58.54
	(15). ii(8)+i(7)	58.04
	(16). ii(11)+i(6)	56.21
	(17). ii(11)+i(7)	56.55
IV	(18). iii(14)+i(5)	59.32
	(19). iii(13)+i(5)	58.70
	(20). iii(15)+iii(16)	58.02

The lists were automatically translated into Bengali using the synset-based English to Bengali bilingual dictionary being developed as a part of the EILMT project. The duplicate removal technique was also applied on the translated synsets to reduce the error. A human translator translated the non-translated entries containing word combinations, idioms etc.

Sense-wise separated word groups give a clue to pattern-based similarity in Bengali to English bilingual dictionary[4]. The sense disambiguation algorithm [10] based on the similarity clue was applied on the translated Bengali synsets.

Two native translators carried out the evaluation. Inter-translator agreement was checked through a statistical measure, *kappa* [6]. The *kappa* coefficient (k) varies within the range from 0.44 to 0.56. It shows a moderate agreement and achieves significant impact on the overall translation process. Further details about the evaluation can be seen in [10].

5.3 Document Classification

Machine learning approaches need a large amount of training data. A good resource-based classifier can negate the need for such large amount of data. In this approach, we aim to evaluate a majority-based emotion classifier that considers *BengWAL*. Based on the terms present in the document, it assigns the probable Ekman emotion classes. The algorithm for resource-based emotion analysis is as follows:

1. For each word in the document, apply stop word removal and stemming.
2. Look up the word in the *BengWAL.*
3. Assign emotion classes to the words based on their presence in any of the six lists of *BengWAL.*
4. Assign the emotion class to a document based on the majority presence of words of that class.

The experiment was conducted with/without stop word removal and with/without stemming. We used an in-house stemmer or suffix stripper and a manually prepared stop word list to accomplish the present task. The overall emotion of a document was identified by the majority of the word level emotions. The evaluation was conducted based on the two best emotion tags according to the number of emotion words of those emotion classes. In Table 5, we present the average *F-Scores* of resource-based emotion classification with respect to all emotion classes.

It is observed that the resource-based system identifies the document level emotions based on the presence of affective words only. The presence of words reflecting indirect emotions, idioms, metaphors and negations are not identified by the system. Moreover the emoticons are the most valuable clues in identifying emotions. But, the resource-based system has not considered the presence of such emoticons. For this reason, the performance of the system has been saturated to 51.07% average *F-Score* for the 110 test documents.

Table 5. F-Scores (in %) of resource based classification

Experiments	F-Score (in %)
BengWAL	46.03
Stop Word Removal	48.28
Stemming	51.07

6 Comparative Results

We have conducted the emotion class-wise evaluation in terms of *F-Scores* that were measured by calculating the average *F-Scores* of the best two tag combinations of the documents. One of the reasons of selecting the best two emotion tags is that a large number of documents in the development set are annotated with single or double emotion tags. It has been observed that the performance of the system in terms of *F-Score* has not improved by adding more than two tags in GSDT and SGDT sets in the extended evaluative experiments.

The machine learning (Word to Sentence and Sentence to Document, i.e., *W-S-S-D*) and resource-based (*BengWAL*) approaches were evaluated against 110 gold standard test documents. Table 6 shows the detailed *F-Scores* of the best two tag combinations per emotion

[4] http://home.uchicago.edu/~cbs2/banglainstruction.html

class obtained using the machine learning (*W-S-S-D)* and resource-based (*BengWAL*) approaches. It is found that the resource-based average *F-Scores* were generally low as compared to the machine learning approach for emotion analysis.

Table 6. Document level emotion tagging

Emotion Class	{GSDT, SGDT}	W-SS-D	Beng WAL
Happy	{d*max1, dmax1´*}	61.23	50.89
	{*dmax1 , dmax2´*}	58.11	48.02
	{*dmax2 , dmax1´*}	57.08	47.94
	{*dmax2 , dmax2´*}	58.56	52.73
	Average	**58.74**	**49.89**
Sad	{d*max1, dmax1´*}	60.98	50.02
	{*dmax1 , dmax2´*}	61.08	51.34
	{*dmax2 , dmax1´*}	59.77	49.20
	{*dmax2 , dmax2´*}	61.32	51.22
	Average	**60.78**	**50.44**
Anger	{d*max1, dmax1´*}	61.57	48.66
	{*dmax1 , dmax2´*}	59.22	49.79
	{*dmax2 , dmax1´*}	59.69	50.24
	{*dmax2 , dmax2´*}	60.54	52.44
	Average	**60.25**	**50.28**
Disgust	{d*max1, dmax1´*}	57.87	47.65
	{*dmax1 , dmax2´*}	58.06	52.68
	{*dmax2 , dmax1´*}	58.17	48.32
	{*dmax2 , dmax2´*}	59.51	49.12
	Average	**58.50**	**49.44**
Fear	{d*max1, dmax1´*}	57.34	52.79
	{*dmax1 , dmax2´*}	57.81	50.80
	{*dmax2 , dmax1´*}	59.37	51.60
	{*dmax2 , dmax2´*}	58.25	52.42
	Average	**58.19**	**51.90**
Surprise	{d*max1, dmax1´*}	60.33	50.31
	{*dmax1 , dmax2´*}	60.81	52.44
	{*dmax2 , dmax1´*}	61.37	48.90
	{*dmax2 , dmax2´*}	60.25	52.93
	Average	**60.69**	**51.14**

These low scores can be attributed to two possible sources of error. The first source is the absence of some sense annotation for the words present in the documents. Though the automatic sense disambiguation was carried out for *BengWAL*, the exact senses of the words are not always available as it is evident in the moderate agreement of *kappa*. The second reason is the coverage of the current *BengWAL*. The generation of the modified *BengWAL* is under construction and hence, not all the synsets have been assigned to the appropriate lists of emotion classes. On the other hand, the emoticons present in the blogs were not considered in this approach. Moreover, the idioms and metaphors used in the blogs were not tagged by the resource-based approach as expected.

Though the resource-based approach did not produce significant results in comparison with *W-S-S-D* approach, this resource-based methodology is simpler and has a less computational cost than *W-S-S-D*. Besides, the machine learning based *W-S-S-D* model produces better result by handling the negative words at the post-processing step [12]. Presently, the inclusion of emoticons in the *BengWAL* and the handling of negative words are being conducted to improve the resource-based approach.

It has been observed that the *F-Scores* that are achieved for the annotated documents with single emotion tag are better in comparison to the *F-Scores* for multiple tags. Hence, the tag combination {d*max1, dmax1´*} performs better for the documents with single emotion tag while the other three combinations perform satisfactorily for the documents with multiple tags.

It is also found that the *happy* emotion tag occurs more frequently with *surprise* emotion tag rather than other emotion tags in the annotated as well as the system-generated outputs. Consequently, *sadness* co-occurs with *angry*, *disgust* and *fear* emotion types rather than *happy* or *surprise*. A closer investigation suggests that the emotion tags with similar emotional slants are present in a sentence with reasonable higher frequencies rather than the emotions with dissimilar emotional slants.

The errors of the machine learning system occur in identifying metaphors, idioms and sentences containing indirect emotional clues. Beside these error cases, the resource-based system also suffers from handling negative words.

7 Conclusions

In the present task, the document level emotion tagging was carried out using two separate approaches, namely, a machine learning approach which assigned emotions to Bengali blog documents based on word to sentence and sentence to document level granularity, and the resource-based approach that incorporated the emotion lexicon, Bengali WordNet Affect Lists (*BengWAL*). To improve the coverage of *BengWAL*, an attempt was made to include a bigger number of emotion words and emoticons along with domain specific knowledge in this lexical resource. The developed document level emotion tagger can be used in an emotion based information retrieval system where retrieved documents will match the user defined query word(s) and emotion specification. The idea of assigning two emotion tags to documents can then be related to the ranking of the retrieved sentences and documents. Emotion analysis related to the effect of metaphors (especially in blogs) is the research area to be explored in future. The clause level analysis of complex emotional sentences may be another area for further studies. Among all concerns, sentiments or emotions of people are important because people's sentiment has great influence on our society. In future, we will attempt to investigate emotions at the entity level (event, topic etc.) [25], since there is a close association among people, topic and emotion.

References

1. **Banea, C., Mihalcea, R. & Wiebe, J. (2008).** A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. *6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco, 2764–2767.
2. **Cardie, C., Wiebe, J., Wilson, T. & Litman, D. (2003).** Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering. *AAAI Spring Symposium on New Directions in Question Answering*, Palo Alto, California, USA, 20–27.
3. **Chaumartin, F. (2007).** Upar7: A knowledge-based system for headline sentiment tagging. *4th International Workshop on Semantic Evaluations* (*SemEval'07)*, Prague, Czech Republic, 422–425.
4. **Chesley, P., Vincent, B., Xu, L. & Srihari, R.K. (2006).** Using verbs and adjectives to automatically classify blog sentiment. *AAAI Symposium on Computational Approaches to Analysing Weblogs* (*AAAI-CAAW 2006),Stanford, CA, USA, 27–29.*
5. **Clore, G. L., Ortony, A. & Foss, M.A. (1987).** The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, 53(4), 751–766.
6. **Cohen, J. (1960).** A coefficient of agreement for nominal scales. *Educational and Psychologica l Measurement*, 20(1), 37–46.
7. **Cortes, C. & Vapnik, V. (1995).** Support-Vector Network. *Machine Learning*, 20(3), 273–297.
8. **Das, D. & Bandyopadhyay, S. (2009).** Emotion Tagging – A Comparative Study on Bengali and English Blogs. *7th International Conference on Natural Language Processing (ICON-2009)*, Hyderabad, India, 177–184.
9. **Das, D. & Bandyopadhyay, S. (2009).**Word to Sentence Level Emotion Tagging for Bengali Blogs. *ACL-IJCNLP 2009 Conference Short Papers,* Suntec, Singapore, 149–152.
10. **Das, D. & Bandyopadhyay, S. (2010).** Developing Bengali WordNet Affect for Analyzing Emotion. *23rd International Conference on Computer Processing of Oriental Languages*, *California, USA,* 35–40.
11. **Das, D. & Bandyopadhyay, S. (2010).** Labeling Emotion in Bengali Blog Corpus – A Fine Grained Tagging at Sentence Level. *8th WorshoponAsianLanguageResources (COLING-2010)*, Beijing, China, 47–55.
12. **Das, D. & Bandyopadhyay, S. (2010).** Sentence Level Emotion Tagging on Blog and News Corpora. *Journal of Intelligent System*, 19(2), 145–162.
13. **Das, D. & Bandyopadhyay, S. (2010).**Sentence to Document Level Emotion Tagging – A Coarse-grained Study on Bengali Blogs. *2nd Mexican Conference on Pattern Recognition: Advances in pattern recognition (MCPR'10),* 332–341.
14. **Das, D., Kolya, A.K., Ekbal, A. & Bandyopadhyay, S. (2011).** Temporal Analysis of Sentiment Events-A Visual Realization and Tracking. *12th International Conference on Intelligent Text Processing and Computational Linguistics, (CICLing-2011), A. Gelbukh (Ed.)*, LNCS 6608, 417–428, Tokyo, Japan.
15. **Denecke, K. (2008).** Using SentiWordNet for multilingual sentiment analysis. IEEE

24th *International Conference on Data Engineering Workshop*, Cancún, México, 507–512.

16. **Ekman, P. (1993).** Facial expression and emotion. *The American Psychologist*, 48(4), 384–392.
17. **Esuli, A. & Sebastiani, F. (2006).** SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *5th Conference on Language Resources and Evaluation*, Genoa, Italy, 417–422.
18. **Grefenstette, G., Qu, Y., Shanahan, J.G. & Evans, D.A. (2004).** *Coupling niche browsers and affect analysis for an opinion mining application. 7th international conference on Computer-Assisted Information Retrieval* (Recherché d'Information*et ses Applications)* RIAO 2004, Avignon, France,186–194.
19. **Katz, P., Singleton, M. & Wicentowski, R. (2007).** Swat-mp: the semeval-2007 systems for task 5 and task 14, *4th International Workshop on Semantic Evaluations (SemEval '07),* Prague, Czech Republic, 308–313.
20. **Kipper-Schuler, K. (2005).** *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.
21. **Ku, L.W., Liang, Y. T. & Chen, H.H. (2006).** Opinion extraction, summarization and tracking in news and blog corpora. *AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs,* California, USA, 100–107.
22. **Lafferty, J.D., McCallum, A. & Pereira, F.C.N. (2001).** Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *18th International Conference on Machine Learning*. MA, USA, 282–289.
23. **Leopold, E. & Kindermann, J. (2002).** Text categorization with support vector machines. How to represent texts in input space? *Machine Learning*, 46(1–3), 423–444.
24. **Lin, K.H.Y., Yang, C. & Chen, H.H. (2007).** What Emotions do News Articles Trigger in Their Readers? *30th annual international ACM SIGIR conference on Research and development in information retrieval*, Amsterdam, The Netherlands, 733-734.
25. **Liu, B. (2009).** The challenge is still the accuracy of sentiment prediction and solving the associated problems. *Keynote talk in 5th Annual Text Analytics Summit*, Boston, USA, June 1–2*.*
26. **Magnini, B. & Cavaglia, G. (2000).** Integrating subject field codes into wordnet, *Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 1413–1418.
27. **Miller, G. A. (1995).** WordNet: a lexical database for English, *Communications of the ACM*, 38 (11), 39–41.
28. **Mishne, G., de Rijke, M., Nicolov N., Salvetti F., Liberman M. & Martin J.H. (2006).** Capturing Global Mood Levels using Blog Posts. *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, California, USA, 145–152.
29. **Mishne, G. (2005).** Experiments with Mood Classification in Blog Posts. *1st Workshop on Stylistic Analysis of Text for Information Access*. SIGIR 2005, Salvador, Brazil, 53–60.
30. **Pang, B. & Lee, L. (2008).** Opinion mining and Sentiment analysis. *Foundations and Trends in Information Retrieval,* 2(1–2), 1–135.
31. **Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985).** *A comprehensive Grammar of the English Language* (2nd Ed.). New York: Longman.
32. **Sebastiani, F. (2002).** Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.
33. **Sood, S. & Vasserman, L. (2009).** ESSE: Exploring Mood on the Web. *Third International AAAI Conference on Weblogs and Social Media (ICWSM) Data Challenge Workshop*, 32–39.
34. **Strapparava, C. & Mihalcea, R. (2007).** SemEval-2007 Task 14: Affective Text. *4th International Workshop on Semantic Evaluations (SemEval '07),* Prague, Czech Republic, 70–74
35. **Strapparava, C. & Valitutti, A. (2004).** Wordnet-affect: an affective extension of wordnet. *4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 1083–1086.
36. **Turney, P.D. (2002).** Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, USA, 417– 424.
37. **Whitelaw, C., Garg, N., & Argamon, S. (2005).** Using appraisal taxonomies for sentiment analysis. *CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*.Bremen, Germany, 625–631.
38. **Wiebe, J., Wilson, T., Bruce, R., Bell, M. & Martin, M. (2004).** Learning Subjective Language. *Computational Linguistics*, 30 (3), 277–308.
39. **Yang, C., Lin, K.H.Y. & Chen, H.H. (2007).** Emotion classification Using Web Blog Corpora.

IEEE / WIC / ACM International Conference on Web Intelligence, Silicon Valley, USA, 275–278.

40. Yang, C., Lin, K.H.Y. & Chen, H.H. (2007). Building Emotion Lexicon from Weblog Corpora. *45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL'07),* Prague, Czech Republic, 133–136.

41. Zhang, Y., Li, Z., Ren, F. & Kuroiwa, S. (2008). A Preliminary Research of Chinese Emotion Classification Model. *International Journal of Computer Science and Network Security*, 8(11), 127–132.

Dipankar Das has been pursuing a Ph.D. degree in Computer Science and Engineering from Jadavpur University, Kolkata, India since July, 2010. He received a Master of Computer Science and Engineering degree from Jadavpur University, Kolkata, India in July, 2009, and a B.Tech degree in Computer Science and Engineering from West Bengal University of Technology, Kolkata, India in 2005. Since July, 2009, he has been working as a Research Engineer in the research project entitled “Cross Lingual Information Access (CLIA) Systems” funded by Department of Information Technology (DIT), Government of India, and since September, 2010 he has been working as a Research Fellow in the India-Japan cooperative program (DST-JST) 2009 research project entitled “Multidisciplinary Research Field on Sentiment Analysis where AI meets Psychology” funded by the Department of Science and Technology (DST), Government of India. His research interests are in the area of Natural Language Processing, Emotion Analysis, Affect Computing, Information Extraction and Language Generation. He is a student member of the IEEE, ACL, HUMAINE.

Sivaji Bandyopadhyay received a Ph.D. degree in Computer Science and Engineering from Jadavpur University, Kolkata, India in 1998 and a Master of Computer Science & Engineering degree and a Bachelor of Computer Science and Engineering degree from the same University in 1987 and 1985, respectively. Since 1989 he has been working as a faculty member in the Department of Computer Science and Engineering, Jadavpur University. He is engaged in several national and international projects such as “Development of Cross Lingual Information Access (CLIA) Phase II”, “Development of English to Indian Languages Machine Translation Systems (EILMT) Phase II” and “Development of Indian Languages to Indian Languages Machine Translation Systems (IL-ILMT) Phase II”, all supported by Department of Information Technology, Government of India, "An Advanced Platform for Question Answering Systems supported by Indo-French Center for the Promotion of Advanced Research (IFCPAR), "Multidisciplinary Research Field on Sentiment Analysis where AI meets Psychology (SAAIP)" as a Strategic India-Japan Cooperative Program-Project and "Answer Validation through Textual Entailment" with Mexico, both these projects being supported by Department of Science and Technology, Government of India. His research interests are in the areas of Natural Language Processing, Machine Learning, Machine Translation, Sentiment Analysis, Question Answering Systems and Information Extraction. He has more than 100 publications in top conferences and journals and has served as a program chair, a workshop chair and a PC member of IJCNLP, NAACL, NLPKE, ICON and others. He is a member of the ACL and AAMT.

Article received on 11/15/2010; accepted 05/06/2011.