Word to Sentence Level Emotion Tagging for Bengali Blogs

Dipankar Das
Department of Computer Science &
Engineering, Jadavpur University, India
`dipankar.dipnil2005@gmail.com`

Sivaji Bandyopadhyay
Department of Computer Science &
Engineering, Jadavpur University, India
`sivaji_cse_ju@yahoo.com`

Abstract

In this paper, emotion analysis on blog texts has been carried out for a less privileged language like Bengali. Ekman's six basic emotion types have been selected for reliable and semi automatic word level annotation. An automatic classifier has been applied for recognizing six basic emotion types for different words in a sentence. Application of different scoring strategies to identify sentence level emotion tag based on the acquired word level emotion constituents have produced satisfactory performance.

1 Introduction

Emotion is a private state that is not open to objective observation or verification. So, the identification of the emotional state of natural language texts is really a challenging issue. Most of the related work has been conducted for English.

The approach in this paper is to assign emotion tags on the Bengali blog sentences with one of the Ekman's (1993) six basic emotion types such as *happiness, sadness, anger, fear, surprise* and *disgust*. The system consists of two phases, machine learning based word level emotion classification followed by assignment of sentence level emotion tags based on the word level constituents using sense based scoring mechanism. The classifier accuracy has been measured through confusion matrix. Corpus based and sense based *tag weights* have been calculated for each of the six emotion tags and then these emotion *tag weights* have been used to identify sentence level emotion tag. The tuned reference ranges selected from the development set have proved effective on the test set.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 briefly describes the resource preparation. Ma-

chine learning based word level emotion tagging system framework and its evaluation results have been discussed in section 4. Section 5 describes the calculation of *tag weights*, sentence level emotion detection process based on the *tag weights*, evaluation strategies and results. Finally section 6 concludes the paper.

2 Related Work

(Mishne *et al.,* 2006) used several supervised and unsupervised machine learning techniques on blog data for comparative evaluation. Importance of verbs and adjectives in identifying emotion has been explained in (Chesley *et al.,* 2006). (Yang *et al.*, 2007) has used Yahoo! Kimo Blog corpora containing emoticons associated with textual keywords to build emotion lexicons. (Chen *et al.,* 2007) has experimented the emotion classification task on web blog corpora using *Support Vector Machine* (SVM) and *Conditional Random Field* (CRF) and the observed results have shown that the CRF classifiers outperform SVM classifiers in case of document level emotion detection.

3 Resource Preparation

Bengali is a less computerized language and there is no existing emotion word list or Senti-WordNet in Bengali. The English *WordNet Affect lists,* (Strapparava et al., 2004) based on Ekman's six basic emotion types have been updated with the synsets retrieved from the English *SentiWordNet* to have adequate number of emotion word entries.

These lists have been converted to Bengali using English to Bengali bilingual dictionary [1]. These six lists have been termed as *Emotion lists*. A Bengali SentiWordNet is being developed by replacing each word entry in the synonymous set of the English SentiWordNet (Esuli *et al.,* 2006)

---

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

by its equivalent Bengali meaning using the same English to Bengali bilingual dictionary.

A knowledge base for the emoticons has been prepared by experts after minutely analyzing the Bengali blog data. Each image link of the emoticon in the raw corpus has been mapped into its corresponding textual entity in the tagged corpus with the proper emotion tags using the knowledge base. The Bengali blog data have been collected from the web blog archive (www.amarblog.com) containing 1300 sentences on 14 different topics and their corresponding user comments have been retrieved.

4 Word Level Emotion Classification

Primarily, the word level annotation has been semi-automatically carried out using Ekman's six basic emotion tags. The assignment of emotion tag to a word has been done based on the type of the *Emotion Word lists* in which that word is present. Other non-emotional words have been tagged with *neutral* type. 1000 sentences have been considered for training of the CRF based word level emotion classification module. Rest 200 and 100 sentences, verified by language experts to perform evaluation have been considered as development and test data respectively.

4.1 Feature Selection and Training

The Conditional Random Field (CRF) (McCallum, 2001) framework has been used for training as well as for the classification of each word of a sentence into the above-mentioned six emotion tags and one *neutral* tag. By manually reviewing the Bengali blog data and different language specific characteristics, 10 active features have been selected heuristically for our classification task. Each feature value is boolean in nature, with discrete value for intensity feature at the word level.

- *POS information:* We are interested with the verb, noun, adjective and adverb words as these are emotion informative constituents. For this feature, total 1300 sentences has been passed through a Bengali part of speech tagger (Ekbal *et al.* 2008) based on Support Vector Machine (SVM) technique. The POS tagger was developed with a tagset of 26 POS tags[2], defined for the Indian languages. The POS tagger has demonstrated an overall accuracy of approximately 90%.

- *First sentence in a topic*: It has been observed that first sentence of the topic generally contains emotion (Roth *et.al.*, 2005).

- *SentiWordNet emotion word*: A word appearing in the SentiWordNet (Bengali) contains an emotion.

- *Reduplication*: The reduplicated words (e.g., *bhallo bhallo* [good good], *khokhono khokhono* [when when] etc.) in Bengali are most likely emotion words.

- *Question words*: It has been observed that the question words generally contribute to the emotion in a sentence.

- *Colloquial / Foreign words*: The colloquial words (e.g., *kshyama* [pardon] etc.) and foreign words (e.g. Thanks, *gossya* [anger] etc.) are highly rich with their emotional contents.

- *Special punctuation symbols*: The symbols (e.g. !, ?, @ etc) appearing at the word / sentence level convey emotions.

- *Quoted sentence*: The sentences especially remarks or direct speech always contain emotion.

- *Negative word*: Negative words such as *na* (no), *noy* (not) etc. reverse the meaning of the emotion in a sentence. Such words are appropriately tagged.

- *Emoticons*: The emoticons and their consecutive occurrences generally contribute as much as real sentiment to the words or sentences that precede or follow it.

Features	Training	Testing
Parts of Speech	432	221
First Sentence	96	13
Word in SentiWordNet	684	157
Reduplication	18	7
Question Words	23	11
Coll. / Foreign Words	35	9
Special Symbols	16	4
Quoted Sentence	22	8
Negative Words	67	27
Emoticons	87	33

Table 1: Frequencies of different features

Different unigram and bi-gram context features (word level as well as POS tag level) and their combination has been generated from the training corpus. The following sentence contains four features (Colloquial word (*khyama*), special

---

[2]http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

symbol (!), quoted sentence and emotion word (ভালো [*happy*])) together and all these four features are important to identify the emotion of this sentence.

ক্ষমা দাও! "তুমি ভালো লোক"
(*khyama*) (*dao*)!   "(*tumi*) (*bhalo*) (lok)"
(Forgive)!   "(you) (good) (person)"

4.2 Evaluation Results of the Word-level Emotion Classification

Evaluation results of the development set have demonstrated an accuracy of 56.45%. Error analysis has been conducted with the help of confusion matrix as shown in Table 2. A close investigation of the evaluation results suggests that the errors are mostly due to the uneven distribution between emotion and non-emotion tags.

Tags	happy	sad	ang	dis	fear	sur	ntrl
happy		0.01	0.05	0.0	0.0	0.0	0.03
sad	0.006		0.02	0.03	0.0	0.0	0.02
ang	0.0	0.03		0.0	0.02	0.0	0.01
dis	0.0	0.0	0.01		0.01	0.0	0.01
fear	0.0	0.0	0.0	0.0		0.0	0.01
sur	0.02	0.007	0.0	0.0	0.0		0.01
ntrl	0.0	0.0	0.0	0.0	0.0	0.0	

Table 2: Confusion matrix for development set

The number of non-emotional or *neutral* type tags is comparatively higher than other emotional tags in a sentence. So, one solution to this unbalanced class distribution is to split the 'non-emotion' (emo_ntrl) class into several subclasses. That is, given a POS tagset *POS*, we generate new emotion classes, 'emo_ntrl-C'|C∈*POS*. We have 26 sub-classes, which correspond, to non-emotion tags such as 'emo_ntrl-NN' (common noun), 'emo_ntrl-VFM' (verb finite main) etc. Evaluation results of the system with the inclusion of this class splitting technique have shown the accuracies of 64.65% and 66.74% on the development and test data respectively.

5 Sentence Level Emotion Tagging

This module has been developed to identify sentence level emotion tags based on the word level emotion tags.

5.1 Calculation of Emotion Tag weights

Sense_Tag_Weight (**STW**): The *tag weight* has been calculated using *SentiWordNet*. We have selected the basic six words "*happy*", "*sad*", "*anger*", "*disgust*", "*fear*" "*surprise*" as the *seed words* corresponding to each emotion type. The

positive and *negative* scores in the English *SentiWordNet* for each synset in which each of these *seed words* appear have been retrieved and the average of the scores has been fixed as the *Sense_Tag_Weight* of that particular emotion tag.

Corpus_Tag_Weight (**CTW**): This *tag weight* for each emotion tag has been calculated based on the frequency of occurrence of an emotion tag with respect to the total number of occurrences of all six types of emotion tags in the annotated corpus.

Tag Types	CTW	STW
emo_happy	0.5112	0.0125
emo_sad	0.2327	(-) 0.1022
emo_ang	0.0959	(-) 0.5
emo_dis	0.1032	(-) 0.075
emo_fear	0.0465	0.0131
emo_sur	0.0371	0.0625
emo_ntrl	0.0	0.0

Table 3: *CTW* and *STW* for each of six emotion tags with neutral tag

5.2 Scoring Techniques

The following two scoring techniques depending on two calculated *tag weights* (in section 5.1) have been adopted for selecting the best possible sentence level emotion tags.

(1) *Sense_Weight_Score* (**SWS**): Each sentence is assigned a *Sense_Weight_Score* (**SWS**) for each emotion tag which is calculated by dividing the total *Sense_Tag_Weight* (**STW**)of all occurrences of an emotion tag in the sentence by the total *Sense_Tag_Weight* (**STW**) of all types of emotion tags present in that sentence. The *Sense_Weight_Score* is calculated as

$$SWS_i = (STW_i * N_i) / (\sum_{j=1\ to\ 7} STW_j * N_j) \mid i \in j$$

where **SWS**$_i$ is the Sentence level *Sense_Weight_Score* for the emotion tag **i** in the sentence and N_i is the number of occurrences of that emotion tag in the sentence. **STW**$_i$ and **STW**$_j$ are the *Sense_Tag_Weights* for the emotion tags **i** and **j** respectively. Each sentence has been assigned with the sentence level emotion tag **SET**$_i$ for which **SWS**$_i$ is highest, i.e.,

$$SET_i = [\max_{i=1\ to\ 6}(SWS_i)].$$

(2) *Corpus_Weight_Score* (**CWS**): This measure is calculated in a similar manner by using the **CTW** of each emotion tag. The corresponding Bengali sentence is assigned with the emotion tag for which the sentence level **CWS** is highest. The scoring mechanism has been considered for verifying any domain related biasness of emotion and their influence in emotion detection process.

5.3 Evaluation Results of Sentence Level Emotion Tagging

Each sentence in the development and test sets have been annotated *with positive* or *negative* or *neutral* valence and with any of the six emotion tags. The **SWS** has been used in identifying valence scores as there is no valence information carried by **CWS**. The sentences for which the total **SWS** produced *positive*, *negative* and zero (0) values have been tagged as *positive, negative* and *neutral* type. Any domain biasness through **CWS** has been re-evaluated through **SWS** also. We have taken the Bengali corpus from comic related background. So, during analysis on the development set, the **CWS** outperforms the **SWS** significantly in identifying *happy, disgust, fear* and *surprise* sentence level emotion tags. The other **SET**s have been identified through **SWS** as the **CWS** for these **SET**s are significantly less than their corresponding **SWS** as shown in Table 5. The knowledge and information of the reference ranges (shown in Table 4) of **SWS** and **CWS** for assigning valence and six other emotion tags, acquired after tuning of development set, have been applied on the test set. The valence and emotion tag assignment process has been evaluated using accuracy measure on test data. The difference in the accuracies for the development and test sets is negligible. It signifies that the best possible reference range for valence and other emotion tags have been selected. Results in Table 5 show that the system has performed satisfactorily for valence identification as well as for sentence level emotion tagging.

Category	Reference Range
Valence (SWS)	0 to 2.35 (+ve), 0 to -0.56 (-ve) and 0.0 neutral)
happy	0.31 to 1 (CWS)
sad	-0.15 to -1.6 (SWS)
angry	-0.5 to -1.9 (SWS)
disgust	0.18 to 1 (CWS)
fear	0.14 to 1.9 (CWS)
surprise	0.15 to 1.76 (CWS)

Table 4: Reference ranges

6 Conclusion

The hierarchical ordering of the word level to sentence level and from sentence level to document level can be considered as the well favored route to track the document level emotional orientation. The handling of negative words and metaphors and their impact in detecting sentence level emotion along with document level analysis are the future areas to be explored.

Category	Development Before CWS	After SWS		Test
Valence	--	49.56	65.43	66.54
happy	54.15	10.33	63.88	64.28
sad	7.66	42.93	64.56	66.42
angry	15.47	53.44	61.48	60.28
disgust	60.13	17.18	70.19	72.18
fear	55.57	11.54	66.04	67.14
surprise	50.25	12.39	65.45	66.45

Table 5: Accuracies (in %) of valence and six emotion tags in development set before and after applying the reference range and in test set

References

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining.*LREC-06*.

Andrew McCallum, Fernando Pereira and John Lafferty. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and labeling Sequence Data. *ISBN*, 282 – 289.

A. Ekbal and S. Bandyopadhyay. 2008. Web-based Bengali News Corpus for Lexicon Development and POS Tagging. *POLIBITS*, 37(2008):20-29. Mexico.

B. Vincent, L. Xu, P. Chesley and R. K. Srhari. 2006. Using verbs and adjectives to automatically classify blog sentiment.*AAAI-CAAW-06*.

Carlo Strapparava, Rada Mihalcea .2007. SemEval-2007 Task 14: Affective Text. *45th Aunual Meeting of ACL*.

C. Yang, K. H.-Y. Lin, and H.-H. Chen. 2007. Building Emotion Lexicon from Weblog Corpora, *45th Annual Meeting of ACL*, pp. 133-136.

C. Yang, K. H.-Y. Lin, and H.-H. Chen.2007. Emotion Classification from Web Blog Corpora, *IEEE/WIC/ACM*, 275-278.

Cecilia Ovesdotter Alm, Dan Roth, Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. *Human Language Technology and EMNLP*, 579-586.Canada.

G. Mishne and M. de Rijke. 2006. Capturing Global Mood Levels using Blog Posts, *AAAI, Spring Symposium on Computational Approaches to Analysing Weblogs*, 145-152.

Paul Ekman. 1993. Facial expression and emotion. *American Psychologist*, 48(4):384–392.