Word and Sentence Level Emotion Analyzation in Telugu Blog and News

Sadanandam. Manchala[1], D. Chandra Mohan[2] and A. Nagesh[3].

[1] Assistant professor of CSE
KUCE&T, Kakatiya University Campus (KU)
Warangal, Andhra Pradesh, India.
sadanb4u@yahoo.co.in.
[2]Assistant professor of CSE
Matrix Institute of Technology&Engg.,
Hyderbadad, Andhra Pradesh, India.
chanduthecm@gmail.com
[3]Assoc. Professor of CSE
Mahatma Gandhi Institute of Technology
Hyderabad , Andhra Pradesh, India.
akkanagesh@rediffmail.com

Abstract:

*Emotion analysis, a recent sub discipline at the crossroads of information retrieval and computational linguistics is becoming increasingly important from application viewpoints of affective computing.Emotion is crucial to identify as it is not open to any objective observation or verification. In this paper, emotion analysis on blog texts has been carried out for a less privileged language, Telugu and the same system has been applied on the English SemEval 2007 affect sensing corpus containing only news headlines. A set of six emotion tags, namely, happy (*ఆనందము*), sad (*దుఃఖము*), anger (*కోపము*), fear (*భయము*), surprise (*ఆశ్చర్యము*)and disgust (*అసహ్యము*), have been selected towards this emotion detection task for reliable and semi-automatic annotation of blog and news data. Conditional Random Field (CRF) based classifier has been applied for recognizing six basic emotion tags for different words of a sentence. The classifier accuracy has been improved by arranging an equal distribution of emotional tags and non-emotional tag. A score based technique has been adopted to calculate and assign tag weights to each of the six emotion tags. A sense based scoring strategy has been applied to identify sentence level emotion scores for the six emotion tags based on the acquired word level emotion tags. Sentence level emotion tagging has been carried out based on the maximum obtained sentence level emotion scores. Evaluation has been conducted for each emotion class separately on 200 test sentences from each of the Telugu blog and English news data. The system has resulted accuracies of 69.82% and 71.06% for happy, 70.24% and 66.42% for sad, 65.73% and 64.27% for anger, 76.01% and 69.90% for disgust, 72.19% and 73.59% for fear and 70.54% and 66.64% for surprise emotion classes on blog and news test data respectively.*

Keywords:

SentiWordNet, Blog, News, Emotion, CRF, Emotion tag, happy, sad, anger, fear, disgust, surprise.

DOI : 10.5121/ijcsea.2012.2317

1. Introduction

Text does not only contain informative objective contents, but also attitudinal private and subjective information, including emotional states. Human emotion described in texts is an important cue for our communication. So, the identification of the emotional state from texts is really a challenging issue. Recently an increasing amount of research has been devoted to investigate the ways of recognizing favorable and unfavorable sentiments towards specific subjects within natural language texts [7].

Emotion classification is a recent sub discipline at the crossroads of information retrieval and computational linguistics. Emotions may be expressed by a single word or a group of words. Sentence level emotion identification process plays an important role to track emotions or to find out the cues for generating such emotions or to properly identify it. Sentences are the basic information units of any document. The overall document level emotion identification process depends on the emotion expressed by the individual sentences of that document which in turn is based on the emotions expressed by the individual words. Obviously, some pragmatic and discourse level analysis are also necessary to track the actual context of different emotional shades inscribed in the text.

The recent online chat system such as blog can be termed as a popular, communicative and informative repository of text based emotional contents. Emotion Analysis has a rich set of applications, ranging from tracking users' emotion about products or events or about politics as expressed in online forums, to customer relationship management. Question Answering (QA) systems and modern Information Retrieval (IR) systems are increasingly incorporating emotion analysis within their scope. Not only QA and IE, but a wide range of other Natural Language Processing tasks have started to use emotional information.

In this paper, the investigation is mainly focused on the emotion analysis of the Telugu blog texts and news headlines of English SemEval 2007 affect sensing corpus [1] according to Ekman's (1993) [2] six basic emotion types such as *happiness (*ఆనందం*)*, *sadness (*దుఃఖం*)*, *anger (*కోపం*)*, *fear (*భయం*)*, *surprise (*ఆశ్చర్యం*)*and *disgust (*అసహ్యం*)*. The non-emotional words are tagged with *neutral* type. The word level annotation in both the corpus has been carried out semi automatically. Linguistic verification of the annotated data has been found to be satisfactory on both blog and news data. The emoticons that appear in the Telugu blog texts have been assigned with their proper tags using a predefined knowledge base.

The Conditional Random Field (CRF) based machine learning approach [5] has been used for word level emotion classification. The emotion classification system has demonstrated accuracies of 71.89% for Telugu and 80.91% for English on 300 sentences of the development sets for both the languages. Error analysis has been incorporated and equal distribution of emotion tags with the non-emotion tag has been organized to improve word level emotion tagging accuracy of the classifier. 76.73% and 87.65% overall accuracies have been achieved on 200 test sentences for Telugu blog and English news data respectively.

Each test sentence of the Telugu blog data has been annotated with single emotion tag and the whole test set has been verified by language experts. The English test sentences have been separated into the emotion classes corresponding to the possible emotion types that contain annotated scores in the range fifty to hundred [50-100] in the news headlines. Evaluations for both blog and news have been carried out separately for each of the six emotion classes. Six different emotion tag weights have been calculated and these emotion tag weights have been applied on the word level emotion tagged data to acquire sentence level emotion scores for each emotion type. A sentence level emotion tag that has the maximum emotion score has been assigned to each sentence. The sentences have been classified and separated into six emotion classes accordingly. The Telugu blog and the English news data, each containing 200 test sentences have shown accuracies of 69.82% and 71.06% for *happy*, 70.24% and 66.42% for *sad*, 65.73% and 64.27% for *anger*, 76.01% and 69.90% for *disgust*, 72.19% and 73.59% for *fear* and 70.54% and 66.64% for *surprise*emotion classes respectively.

The rest of the paper is organized as follows. Section 2 describes the related works done in this area. Preparation of relevant resources has been described in section 3. The description of the CRF-based word level emotion tagging system framework and its evaluation are specified in section 4. Section 5 describes the method of calculating six emotion tag weights for assigning sentence level emotion tag and post processing for negative words handling. Emotion class wise evaluation mechanisms and associated results are discussed in section 6. Finally section 7 concludes the paper.

2. Related work

Different unsupervised, supervised and semi supervised strategies have been adopted for decades to identify and classify emotions. Characterization of words and phrases according to their emotive tone has been described in [6]. The system classifies the reviews into two types: *recommended* (thumbs up) and *not recommended* (thumbs down) using the *semantic orientation* of the phrases in the review. But in many domains of text, the values of individual phrases may bear little relation to the overall sentiment expressed by the text. In the article [7], the emotion analysis task is defined analogous to topic classification and underscores the difference between the machine learning methods and human produced baseline model. A short study described in [8] addresses the text categorization task using machine learning techniques. Several supervised and unsupervised machine learning classification techniques on blog data and comparative evaluation are described in [9]. *Support Vector Machine* (SVM) has been used in this work to identify the intensity of the community mood.

Importance of verbs and adjectives in identifying emotion has been explained in [10]. Each post from a blog has been classified as *objective*, *subjective-positive*, or *subjective-negative*. This approach is topic and genre independent. The annotation of opinionated content of a language has been described in [11].

A Yahoo! Kimo Blog corpus has been used in [18] to build emotion lexicons. In their studies, emoticons were used to identify emotions associated with textual keywords. A system for classifying news articles according to the readers' emotions instead of the authors' has been implemented in [19]. Experiments have been carried out for the emotion classification task on

web blog corpora using SVM and CRF machine learning techniques. It has been observed that the CRF classifiers outperform SVM classifiers in case of document level emotion detection.
The text-based emotion prediction problem using supervised machine learning with the *SNoW*learning architecture has been discussed in [14]. An affective text task on news headlines at SemEval 2007 for emotion and valence level identification has been described in [15]. The assessment of textual affects using real world knowledge has been described in [16]. Opinions mining at word, sentence and document levels from news and web blog articles and opinion summarization have been described in [17].

Most of the works have been carried out for English. Telugu is less privileged and less computerized than English. As far our knowledge goes, no work on Emotion analysis has yet been initiated for Telugu. So, we have focused on Telugu. For evaluative measures, the system has been applied on the English data. Most of the above discussed machine learning based models have considered sentence as their basic key constituent whereas this present work deals with word for fine grained analysis. The calculation of emotion tag weights and corresponding emotion scores for each sentence are the key contributions of this present work.

3. Resource Preparation

The sentiment lexicon, *SentiWordNet* [4] and emotion word lists like *WordNet Affect lists* [3] are available in English and these resources have been utilized for emotion analysis task for English news data. But, Telugu is a less computerized language and there is no existing emotion lexicon in Telugu. To overcome this problem, a Telugu*SentiWordNet* is being developed by replacing each word entry in the synonymous set of the English *SentiWordNet* by its possible Telugu meaning using English to Telugu bilingual dictionary (Oxford English to Telugu Dictionary) and Google translator. This modified *SentiWordNet* for Telugu has been considered as emotion based lexicon throughout the work.

A knowledge base (shown in ***Table 1***) for the emoticons has been prepared by experts after minutely analyzing the Telugu blog data. Each image link of the emoticon in the raw corpus has been mapped into its corresponding textual entity in the tagged corpus according to their proper emotion types using the knowledge base. It has also been used during semi-automatic annotation of Telugu blog data.

Table 1. Knowledgebase for tracking emoticons during annotation

Symbol	Emoticons	Source Tag	Type
	:^)	*<emo_icon_happy>*	*happy*
	**:-@**	*<emo_icon_anger>*	*anger*
	:-$	*<emo_icon_disgust>*	*disgust*
	:'(	*<emo_icon_fear>*	*fear*
	:(	*<emo_icon_sad>*	*sad*
	°o°	*<emo_icon_surprise>*	*surprise*
	:|	*<emo_icon_neutral>*	*neutral*

3.1 WordNet Affect

The English *WordNet Affect*, based on Ekman's (1993) [7] six emotion types (*joy, fear, anger, sadness, disgust, surprise*) is a small lexical resource compared to the complete *WordNet* but its affective annotation helps in emotion analysis. Some collection of *WordNet Affect* synsets was provided as a resource for the shared task of *Affective Text* in *SemEval-2007* [15]. The whole data is provided in six files named by the six emotions. Each file contains a list of synsets and one synset per line. An example synset entry from *WordNet Affect* is shown as follows.

a#00117872 angered enraged furious infuriated maddened

The first letter of each line indicates the part of speech (POS) and is followed by the *affectID*. The representation was simple and easy for further processing. We have retrieved and linked the compatible synsetID from the recent version of *WordNet 3.0* with the *affectID* of the *WordNet Affect* synsets using an open source tool[1]. The linking of the *WordNet Affect* synsets with their corresponding synsets of *WordNet 3.0* has been done (*n#05587878* ←→ *07516354-n anger choler ire*). The differences between emotions, cognitive states and affects are not analyzed in the present work. Our main focus in the task was to develop an equivalent resource in Telugu for analyzing emotions.

3.2 Expansion of WordNet Affect using SentiWordNet

It has been observed that the *WordNet Affect* contains fewer number of emotion word entries. The six lists provided in the *SemEval 2007* shared task contain only 612 synsets in total with 1536 words. The detail distribution of the emotion words as well as the synsets in six different lists

[1] http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=21&Itemid=59

according to their POS is shown in ***Table 2(a), Table2(b).*** Hence, we have expanded the lists with adequate number of emotion words using *SentiWordNet 3.0* before attempting any translation of the lists into Telugu. *SentiWordNet* assigns each synset of *WordNet* with two coarse grained subjective scores such as *positive*, *negative* along with an *objective* score. *SentiWordNet* contains more number of coarse grained emotional words than *WordNet Affect*. We assumed that the translation of the coarse grained emotional words into Telugu might contain more or less fine-grained emotion words. Our aim is to increase the number of emotion words in the *WordNet Affect* using *SentiWordNet*. Both of the two resources are developed from the *WordNet*. Hence, each word of the *WordNet Affect* is replaced by the equivalent synsets retrieved from *SentiWordNet* if the synset contains that emotion word. The POS information in the *WordNet Affect* is kept unchanged during expansion. The POS of a *SentiWordNet* synset entry is followed by a *synset ID*, *positive* and *negative* scores and the synset contains sentiment words. The distributions of expanded synsets for each of the six emotion classes based on four different POS types (*Noun*, *Verb*, *Adjective* and *Adverb*) are shown in Table 1. But, we have kept the duplicate entries at synset level for identifying the emotion related scores in our future attempts by utilizing the already associated *positive* and *negative* scores of *SentiWordNet*. The percentage of entries in the updated word lists are increased by 69.77 and 74.60 at synset and word levels respectively.

Emotion Classes	*WordNet Affect* **Synset (*S*)** [***After SentiWordNet updating***]			
	Noun	***Verb***	***Adjective***	***Adverb***
	S	***S***	***S***	***S***
Anger	48**[*198*]**	19[***103***]	39[***89***]	21 [***23***]
Disgust	3 [***17***]	6 [***21***]	6[***38***]	4 [***5***]
Fear	23[***89***]	15 [***48***]	29[***62***]	15 [***21***]
Joy	73[***375***]	40[***252***]	84[***194***]	30 [***45***]
Sadness	32[***115***]	10 [***43***]	55 [***129***]	26 [***26***]
Surprise	5 [***31***]	7 [***42***]	12[***33***]	4 [***6***]

Table 2(a). Number of POS based Synsets in six *WordNet Affect* lists before and after updating using *SentiWordNet*

Emotion Classes	*WordNet Affect* **Word (W)** [***After SentiWordNet updating***]			
	Noun	***Verb***	***Adjective***	***Adverb***
	W	***W***	***W***	***W***
Anger	99 [***403***]	64[***399***]	120[***328***]	35 [***50***]
Disgust	6 [***21***]	22 [***62***]	34 [***230***]	10 [***19***]
Fear	45 [***224***]	40[***243***]	97 [***261***]	26 [***49***]
Joy	149[***761***]	122[***727***	203[***616***]	65[***133***]
Sadness	64 [***180***]	33 [***92***]	169[***779***]	43 [***47***]
Surprise	8 [***28***]	28 [***205***]	41 [***164***]	13 [***28***]

Table 2(a). Number of POS based Words in six *WordNet Affect* lists before and after updating using *SentiWordNet*

The Telugu blog data has been collected from the web blog. The 14 different topics and their corresponding user comments containing 1500 sentences have been retrieved. Each of 1253

sentences, tagged with scores for all six emotion tags have been retrieved from the English SemEval 2007 corpus [1].

3.3 Translation of Expanded WordNet Affect into Telugu

Telugu WordNet affect lists are shown in ***Table 3***. We have mapped the *affectID* of the *WordNet Affect* to the corresponding *synsetID* of the *WordNet 3.0*. This mapping helps to expand the *WordNet Affect* with the recent version of *SentiWordNet* 3.0. As there is no Telugu *WordNet* is freely available and it is being developed based on the English *WordNet*, the synsets of the expanded lists are automatically translated into Telugu equivalent synsets based on the *synsetIDs*usingopen source google dictionary API and the English to Telugu bilingual dictionary. The number of translated Telugu words and synsets for six affect lists are shown in ***Table 3***. There are some translated samples that contain word level as well as phrase level translations

(e.g., గట్టిగా ఉద్దటంవల్లబ ధకలుగు (*gattigarudhatamvallabadhakalugu*) '*chafe*', ప్రత్యహించెండుకొత్తెచప్పట్లు (*prostyahinchendukykottechappatlu*) '*cheer*', మనసుకుఅసవ్కర్యమైనస్థితి (*manasukuasavkaryamynasthithi*) '*dysphoria*', etc.).

Emotion Classes	**Telugu *WordNet Affect* list**		
	***Translated* (#Words)**	***Non-Translated* (#Words)**	***Translated* (#*Synset)***
Anger	240	205	1033
Disgust	22	69	218
Fear	80	210	615
Joy	379	387	2940
Sadness	133	299	846
Surprise	74	34	456

Table 3: Number of translated and non-translated Synset and word entries in six Telugu WordNet Affect lists.

4. Word Level Emotion Tagging System

Primarily, the word level annotation has been carried out semi-automatically. The manual assignment of emotion tag to a word has been done with the help of the *Emotion list* in which that word is present. Other non-emotional words have been tagged with *neutral* type. The word level emotion annotated sentences have been verified by language experts. Total 1500 Telugu blog sentences and 1253 English news headlines have been annotated with six emotion tags at word level. Out of 1500 annotated Telugu blog sentences, 1000 sentences and out of 1253 annotated English news sentences, 753 sentences have been considered for training with CRF based word

level emotion tagging system. Out of the rest 500 sentences of Telugu blog and English news data, 300 and 200 sentences have been used as development and test set respectively.

Here, we have used CRF classifier for classifying emotion and non-emotion words into their appropriate classes and tag them with their proper emotion and *neutral* tags. The CRF classifier performs the classification task at sentence level, and thus it carries out word level classification task without any loss of emotional constituents at sentence level.

4.1 Feature Selection & Training

Feature plays a crucial rule in the CRF framework. By manually reviewing the Telugu blog data and English news corpus and their different language specific characteristics, 10 active features have been selected heuristically for our classification task. Each feature 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 1500 blog sentences have been passed through a Telugu part of speech tagger based on SVM framework. The POS tagger was developed with a tag set of 27 POS tags (http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf), defined for the Indian languages. The POS tagger has demonstrated an overall accuracy of approximately 90%. The 1253 sentences, collected from the English SemEval 2007 Affect Sensing Corpus, has been POS tagged with an open source Stanford Maximum Entropy based POS tagger [18]. The best reported accuracy for the POS tagger on the Penn Treebank is 96.86% overall and 86.91% on previously unseen words.

First sentence in a topic: It has been observed that first sentence of the topic generally contains emotion [14]. But during sentence level emotion analysis, all English news headlines considered for this work are equally important. So, this feature has been identified as less significant during feature level analysis on the English development set.

SentiWordNet emotion word**:** A word appearing in the SentiWordNet (English and Telugu) generally contains emotion. For word level emotion classification, it is necessary to disambiguate emotion and non-emotion words properly. This feature helps the classifier to clearly define emotion and non-emotion words.

Reduplication: The reduplicated words (e.g.,బాబాబా బాబాబా [good good] ఆ ఆ ఆ ఆ [many many] etc.) in Telugu are most likely emotion words. English reduplicated words ([so so] etc.) have also been taken into consideration as they are emotion words.

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

Special punctuation symbols: The symbols (e.g., !, ?, @ etc) appearing at the word / sentence level convey emotions. Such symbols are appropriately tagged. The number of occurrences of special punctuation symbols attached to a word is an important word level emotion feature.

Quoted sentence: The sentences especially remarks or direct speech always contain emotion.
Length of a Sentence: Sentence length is a crucial factor for emotion classification task. Sentences containing maximum of eight (8) words have demonstrated a satisfied word level emotion tagging accuracy.

Emoticons: The emoticons generally contribute as much as real sentiment to the words that precede or follow it. The consecutive occurrence of such emoticons emphasizes the preceding word or sentence.

ಮನ್ನಿಂಚು! “ಮೀರು/ನೀವು ಮಂಚಿ ವ್ಯಕ್ತಿ”

(Manninchu)!“(Meeru)/(neevu) (manchi) (vyakthi)
(Forgive)! “(you) / (you) (good) (person)”

Different unigram and bi-gram context features (word level as well as POS tag level) and their combinations have been generated from the training corpus. The above Telugu blog sentence contains four features (Colloquial word (*manninchu*), special symbol (!), quoted sentence and emotion word (ಆನಂದ[*happy*])) together and all these four features are important to identify the emotion of this sentence.

4.2 Evaluation

It has been observed during the development phase that certain features (e.g. First sentence in a topic) have played no role in word level emotion tagging process for English news sentences. This is a specific phenomenon for the SemEval 2007 data as all sentences in the corpus are news headlines and hence are always first sentences in the corresponding text document. The inclusion of such feature has not improved the system accuracy and has been removed from the feature set.

The news headlines do not contain emoticons and this feature has therefore been excluded from the active feature set. The POS feature (only *adjective*, *noun*, *verb* and *adverb* words) has played an important role in this task. Special Symbols and their number of occurrences have been selected as one of the features and this feature has enhanced the performance of the system by around 3% for blog and news data respectively. During CRF-based training phase, current token word with two previous and next words and their corresponding POS have been selected as context feature for that word. The frequencies of different features in training and test set have been shown in ***Table 4***. Evaluation results of the development set of 300 sentences have demonstrated an accuracy of 71.89% for Telugu blog data and 80.91% for English news data.

Feature Type	Frequency in the Corpus			
	Blogs		News	
	Training	Test	Training	Test
Part of Speech	2132	678	1647	532
First Sentence	96	13	753	200
Word in SWN	684	157	1157	175
Reduplication	18	7	5	0
Question Words	13	11	23	9
Colloquial/foreign Words	35	9	8	0
Special Symbols	65	43	17	8
Quoted Sentence	22	8	7	3
Length of Sentence(>=8)	732	239	589	110
Emotions	87	33	0	0

Table 4. Frequencies of different features in the Trainingand test corpus of Telugu blogand English news data.

Error analysis has been conducted with the help of confusion matrices as shown in ***Table 5*** and ***Table 6*** for Telugu blog data and English news headlines data respectively. A close investigation of the evaluation results suggest that the errors are mostly due to the uneven distribution between emotion and non-emotion tags. The number of non-emotional or neutral type tags is comparatively higher than that of 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 effectively. That is, given a POS tagset*POS*, we generate new emotion classes, 'emo_ntrl-C'|C ***∈*** *POS*. We have twenty seven (27) subclasses in the Telugu POS tagset and forty five (45) subclasses in the English POS tagset, which correspond, to non-emotion regions such as 'emo_ntrl-NN' (common noun), 'emo_ntrl-VFM' (verb finite main), 'emo_ntrl-JJ' (adjective) etc. Evaluation results of the system have shown the improved accuracies of 72.85% for blog and 87.65% for news with the inclusion of this class splitting technique that has been applied on 200 test sentences from each category. As Telugu is a morphologically rich language and it takes variety of suffixes, the word level emotion tagging accuracy that has been achieved is less in comparison with English.

Tag	anger	dis	fear	hpy	sad	sur	neutral
anger	-	0.0	0.02	0.0	0.03	0.0	0.01
disgust	0.01	-	0.01	0.0	0.0	0.0	0.01
fear	0.0	0.0	-	0.0	0.0	0.0	0.01
happy	0.05	0.0	0.0	-	0.01	0.0	0.03
sad	0.02	0.03	0.0	0.006	-	0.0	0.02
surprise	0.0	0.0	0.0	0.02	0.007	-	0.12
neutral	0.0	0.0	0.0	0.0	0.0	0.0	-

Table 5. Confusion matrix for the development set of Telugu Blog.

Tag	anger	dis	fear	hpy	sad	sur	neutral
anger	-	0.0	0.02	0.0	0.03	0.0	0.01
disgust	0.01	-	0.01	0.0	0.0	0.0	0.01
fear	0.0	0.0	-	0.0	0.0	0.0	0.01
happy	0.05	0.0	0.0	-	0.01	0.0	0.03
sad	0.02	0.03	0.0	0.006	-	0.0	0.02
surprise	0.0	0.0	0.0	0.02	0.007	-	0.12
neutral	0.0	0.0	0.0	0.0	0.0	0.0	-

Table 6. Confusion matrix for the development set of English News headlines

5 . Sentence Level Emotion Tagging

This module has been developed to identify sentence level emotion tags based on the word level emotion tags (assigned by the CRF based Emotion tagger) for each sentence of test set for each of the languages. Before that, we have calculated sense based *tag weights* for each of the six emotion tags.

Sense_Tag_Weight(***STW***): This 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 type of emotion tag. 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. ***Table 7*** shows the values of ***STW*** of six emotion tags. Neutral tag has been assigned with ***STW*** as zero.

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 the emotion tag in the sentence by the total *Sense_Tag_Weight*(***STW***) of all types of emotion tags present in that sentence.

Tag Type	**Sense_Tag_Weight**
emo_anger	0.0125
emo_disguest	(-) 0.1022
emo_fear	(-)0.5
emo_happy	(-) 0.075
emo_sad	0.0131
emo_surprise	0.0131
emo_neutral	0.0

Table 7.*Sense_Tag_Weight*for each of six emotion tags

Thus, ***SWSi = (STWi * Ni) / (Σ j=1 to 7 STWj * Nj) | i Є j*** where ***SWSi***is the Sentence level *Sense_Weight_Score*for the emotion tag ***i***in the sentence and ***Ni*** is the number of occurrences of that emotion tag in the sentence. Each sentence has been assigned with the sentence level emotion tag ***SET*** for which ***SWSi***is highest.

SET = [**max** i=1 to 7(***SWSi***)].

The sentences have been tagged as *neutral* type if for all emotion tags i, ***SWSi***produced zero (0) emotion score.

5.1 Post processing Strategies

The presence of negative words and their number of occurrences are significant in assigning the final emotion tag for a sentence. We have implemented a rule based post processing module for handling negative words.

The consecutive occurrence of negative words does not reverse the assigned emotion type whereas presence of a single negative word changes the actual emotion type in the completely opposite direction. For example, the following English News sentence

*Smoking **No**Longer Tres Chic in France.*

ಮಾರು ಈ ವ್ಯಾಕುಲತೆ ಚಿಂತೆಯನ್ನು | (Telugu Blog)

(meeru) (vyakulatha)(chendha)(vadhu)
Youworrydo not

has been tagged as "*sad*" by the system but in the gold standard SemEval 2007 news corpus, the emotion scores for "*happy*" and "*surprise* " have been assigned for this sentence.

Considering the single occurrence of the negative word “NO” in the English sentence and “కాదు” (*vadhu*)(not) in the Telugu sentence during post processing phase, the emotion tag of both the sentences have been reversed to “*happy*”, the desired emotion tags for these sentences.
In the following sentence, two consecutive occurrences of negative words (“NO” and “NOT”) do not change the actual emotion expressed by the sentence.

Seduced by Snacks? ***No***, ***Not*** *You.*(English News)

In this case, the system has assigned the “*fear*” tag which only has the significant emotion score in comparison to other emotion scores in the annotated English news corpus. We have applied the rule that two consecutive negative words present in a sentence do not change the sentence level emotion tag. The same rule has been applied for the following Telugu sentence and the annotated “*happy*” emotion tag has also been matched with the system generated “*happy*” emotion tag.

లేదులేదునేను ఇప్పుడు బాగానున్నాను |(Telugu Blog)

(ledhu)(ledhu)(nenu) (eppudu)(baga)(unnanu)
No no I now well am

6. Emotion Class Wise Evaluation

Each Telugu blog sentence in the test set has been annotated with a single emotion tag and the whole test set has been verified by a linguist. The 200 test sentences have been classified into six different emotion classes according to their annotated sentential emotion tag. The test sentences of each emotion class have been passed through the system for assigning a single emotion tag. System generated emotion tag of each test sentence has been compared against its annotated emotion tag. Then the emotion class wise accuracies have been measured by counting the number of sentences whose system assigned emotion tag matches with the emotion class.

Each news headlines in the SemEval 2007 affect sensing corpus has been annotated with individual score for each of the six emotion types. There is no single emotion tag assigned for a sentence. The annotated emotion scores are in the range zero to hundred [0-100] in the English news corpus. We have extracted the sentences that have emotion scores in the range fifty to hundred [50-100] for any emotion tag and each of the sentences is tagged with the corresponding emotion tag. We have extracted 200 test sentences in this manner as our test set. These 200 test sentences have been classified into six emotion classes. Out of 200 test sentences 13 sentences have been classified into more than one emotion class as these sentences are assigned with more than one type of emotion score within the range fifty to hundred [50-100] in the annotated English news SemEval 2007 affect sensing corpus.

Emotion Class	Blogs (Total#Sentences)	News (Total#Sentences)
anger	65.73% (42)	64.27% (56)
disgust	76.01% (31)	69.90% (43)
fear	72.19% (33)	73.59% (48)
happy	69.82% (40)	71.06% (33)
sadness	70.24% (31)	66.42% (51)
surprise	70.54% (23)	66.64% (39)

Table 8. Test set accuracies and total test sentences per emotion class for Blogs and News

The sentences of each emotion class have been passed through the system and accuracies have been measured by counting the number of sentences for which the annotated emotion tag and system generated sentential emotion tag match with respect to the total number of sentences in the class. Test set accuracies for both the Telugu blog data and the English news data for each emotion class have been shown in ***Table 8***.

Results show that the system has performed satisfactorily for both languages although there is a scope for improving the accuracy values. The loss in accuracies has occurred due to the frequent use of metaphoric words in blogs and news as the metaphors are hard to tag with their emotional senses. Apart from these facts, the size of Telugu blog data for word level training is not comparable to the English news data. Again, the word level emotion tagging accuracies have been calculated with respect to all emotion classes whereas the sentence level emotion tagging has been measured with respect to each and individual emotion class.

7. Conclusion

An emotion tagging system for Telugu blog data and English news data that works at the word and the sentence level have been described in this work. The system has demonstrated an overall satisfactory performance throughout the task. The handling of metaphors and their impact in detecting sentence level emotion have not been studied in this work. The phrase level analysis of the input text is the future demand of this system to cope up with the context level discrepancies and fine tuned negation handling. The system can be used in an emotion based information retrieval system where retrieved documents will match the user defined query word(s) and emotion specification. Sometimes, users of the blogs comment on others comments. The identification of such overlapped comments on a given topic is crucial for detecting emotion. Also the tracking of a single user's comments on the same topic as well as on different topics is really important to take care of. These works along with document level analysis are the future areas to be explored. Other future tasks include the application of Emotion tagging system for English blogs.

8. References

[1] Carlo Strapparava, RadaMihalcea.: SemEval-2007 Task 14: Affective Text.Proceedings of the 45th Aunual Meeting of Association for Computational linguistics (2007)

[2] Paul Ekman.: Facial expression and emotion. vol. 48(4), pp. 384--392. American Psychologist (1993)

[3] Carlo Strapparava and Alessandro Valitutti.: WordNet-Affect: an affective extension of WordNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 1083--1086 (2004)

[4] Andrea Esuli and FabrizioSebastiani.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. LREC-06 (2006)

[5] Andrew McCallum, Fernando Pereira and John Lafferty.: Conditional Random Fields: Probabilistic Models for Segmenting and labeling Sequence Data. ISBN, pp. 282 -- 289 (2001)

[6] Peter D. Turney.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 417-- 424 (2002)

[7] Bo Pang and Lillian Lee and ShivakumarVaithyanathan.: Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79--86 (2002)

[8] FabrizioSebastiani.: Machine learning in automated text categorization. ACM Computing Surveys, vol. 34(1) (2002)

[9] G. Mishne and M. de Rijke.: Capturing Global Mood Levels using Blog Posts, Proceedings of AAAI, Spring Symposium on Computational Approaches to Analysing Weblogs, 145--152 (2006)

[10] B. Vincent, L. Xu, P. Chesley and R. K. Srhari.: Using verbs and adjectives to automatically classify blog sentiment. In Proceedings of AAAI-CAAW-06, the Spring Symposia, (2006)

[11] Claire Cardie, JanyceWiebe and Theresa Wilson.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, vol. 39(2), pp. 165--210 (2005)

[12] Changhua Yang, Kevin Hsin-Yih Lin and Hsin-HsiChen.: Emotion Classification Using Web Blog Corpora. IEEE, WIC, ACM International Conference on Web Intelligence, pp. 275--278 (2007)

[13] K. H.-Y. Lin, C. Yang and H.-H.Chen.: What Emotions News Articles Trigger in Their Readers?. Proceedings of SIGIR, pp. 733-734 (2007)

[14] Cecilia OvesdotterAlm, Dan Roth, RichardSproat.: Emotions from text: machine learning for text-based emotion prediction. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing Vancouver, pp. 579 -- 586, British Columbia, Canada (2005)

[15] François-RégisChaumartin.: UPAR7: A knowledge-based system for headline sentiment tagging. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Association for Computational Linguistics, pp. 422--425, (2007)

16] Hugo Liu, Henry Lieberman, Ted Selker.: A Model of Textual Affect Sensing using Real-World Knowledge. IUI '03: Proceedings of the 8th international conference on intelligent user interfaces, ACM, (2003)

[17] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen.: Opinion extraction, summarization and tracking in news and blog corpora. In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, volume AAAI Technical Report, pp. 100--107,March (2006a)

[18] Christopher D. Manning and Kristina Toutanova.: Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC) (2000)

[19] C. Yang, K. H.-Y. Lin, and H.-H. Chen.: Building Emotion Lexicon from Weblog Corpora. Proceedings of the 45th Annual Meeting of ACL, pp. 133--136, (2007)