Computational Linguistics and Chinese Language Processing
Vol. 18, No. 1, March 2013, pp. 79-98

# Emotion Co-referencing - Emotional Expression, Holder, and Topic

**Dipankar Das[*], and Sivaji Bandyopadhyay[+]**

**Abstract**

The present approach aims to identify the *emotional expression*, *holder*, *topic,* and their co-reference from Bengali blog sentences. Two techniques are employed, one is a rule-based baseline system and the other is a supervised system that consists of different syntactic, semantic, rhetorical, and overlapping features. Different error cases have been resolved using rule-based post processing techniques. The evaluative vectors containing emotional expressions, *holders,* and *topics* are prepared from annotated blog posts as well as from system generated output. The evaluation metric, *Krippendorff's **α**,* achieves agreement scores of 0.53 and 0.67 for the baseline and supervised co-reference classification systems, respectively.

**Keywords:** Emotional Expression, Holder, Topic, Co-reference Agreement.

## 1. Introduction

In psychology and in common use, emotion is an aspect of a person's mental state of being, normally based on or tied to the person's internal (physical) and external (social) sensory feeling (Zhang *et al.*, 2008). Emotions, of course, are not linguistic features. Nevertheless, the most convenient access we have to them is through language (Strapparava & Valitutti, 2004). The identification of emotion expressed in the text with respect to the reader or writer is a challenging task (Yang *et al.*, 2009). A wide range of Natural Language Processing (NLP) tasks, from tracking users' emotion about products/events/politics as expressed in online forums or news to customer relationship management, use emotional information.

Currently, emails, blogs, chat rooms, online forums, and even Twitter are being considered as effective communication substrates to analyze the reaction of emotional

[*] Department of Computer Science & Engineering, National Institute of Technology (NIT), Meghalaya, Shillong 793003, India
E-mail: dipankar.dipnil2005@gmail.com

[+] Department of Computer Science & Engineering, Jadavpur University, West Bengal, Kolkata 700032, India
E-mail: sivaji_cse_ju@yahoo.com

catalysts. A blog is a communicative and informative repository of text-based emotional content in the Web 2.0 (Yang *et al.*, 2007). In particular, blog posts contain instant views, updated views, or influenced views regarding single or multiple topics. Many blogs act as online diaries of the bloggers for reporting the blogger's daily activities and surroundings. Sometimes, the blog posts are annotated by other bloggers. In addition, a large collection of blog data is suitable for any machine learning framework.

It has been observed that three major components are crucial in determining the emotional slants from different perspectives: *Emotional Expression*, *Holder*, and *Topic*. Thus, the determination of the emotion *holder* and *topics* from the text helps us track and distinguish users' emotions separately on the same or different *topics*. *Emotional expression* (word or phrase) is the subjective counterpart that can be expressed by a directly affective word ("John is really *happy* enough") or using some indirect notion ("Dream of music is in their eyes and hearts"). The source or *holder* of an *emotional expression* is the speaker, the writer, or the experiencer (Wiebe *et al.*, 2005). Extraction of the emotion *holder* is important in discriminating between emotions that are viewed from different perspectives (Seki, 2007). By grouping opinion *holders* of different stances on diverse social and political issues, we can gain better understanding of the relationships among countries or among organizations (Kim & Hovy, 2006). *Topic*, however, is the real world object, event, or abstract entity that is the primary subject of the emotion or opinion intended by the *holder* (Stoyanov & Cardie, 2008a). *Topic* depends on the context in which its associated *emotional expression* occurs (Stoyanov & Cardie, 2008b). For example, the following Bengali sentence shows the *emotional expression*, its associated *holder*, and *topic*.

**রাশেদ** বলেছেন আপনার **কবিতাটা** পড়তে গিয়ে তার এই
(*Rashed*) (*bolechen*) (*apnar*) (*kobitata*) (*porte*) (*giye*) (*tar*) (*ei*)
**সুন্দর** **কৌতুকটা** মনে পড়ছিলো।
(*sundar*) (*koutukta*) (*mone*) (*porchilo*).
***Rashed** said that he was remembering this **beautiful comic** while reading your **poem**.*
*Emotional Expression*: সুন্দর কৌতুক (**beautiful comic**), *Holder*: < writer**,** রাশেদ (**Rashed**) >**,** *Topic*: কবিতা (**poem**).

In the above example, along with the *emotional expression* and *topic*, the *writer* of the blog post is also considered as a default *holder* according to our assumption, which is based on the nested source hypothesis (Wiebe *et al.*, 2005). Sometimes, the emotional sentences may or may not contain a direct clue for the *emotional expression*. There are certain example sentences that contain an *emotional expression* without a *holder* (*Tar Abhinoy ta satyoi khub* *akorshoniyo chilo* [His acting was really attractive]). Nevertheless, the sentence contains a *topic* (*Tar Abhinoy* [His acting]). Sometimes, even the *emotional expressions* represent the potential *topics*. For example, the Bengali sentence, "*Ami Ramer **doohkho koshte** kende pheli.*" [*I fall into cry on the **sorrow** of Ram*] contains the text "*doohkho koshte*" [*sorrow*] that is treated as both the *emotional expression* and the *topic*. With the above examples and problems in mind, we hypothesize that the notion of user-topic co-references will facilitate both the manual and automatic identification of emotional views. Presently, we have assumed that the *holder* and *topic* are emotion co-referent if they share the same *emotional expression*.

The present task deals with the identification of users' emotions on different *topics* from an annotated Bengali blog corpus (Das & Bandyopadhyay, 2010a). Each sentence of the corpus is annotated with the emotional components, such as *emotional expression* (word/phrase), intensity (*high*, *general*, and *low*), associated *holder*, *topic*(s), and sentential tag of Ekman's six emotion classes (*anger, disgust, fear, happy*, *sad*, and *surprise*).

In this project, a simple rule-based baseline system is developed for identifying the *emotional expressions*, *holders*, and *topics*. The expressions are identified from shallow parsed sentences using Bengali WordNet Affect lists (Das & Bandyopadhyay, 2010b). A simple part-of-speech (POS) tag-based pattern matching technique is employed to identify the emotion *holders* and *topics* with respect to the *emotional expressions*. The presence of emotion *holders* and *topics* in the immediate neighborhood, shallow chunks that refer to their corresponding *emotional expressions*, gives the co-reference clues for the baseline system. The co-reference among the *emotional expressions*, *holders*, and *topics* is measured using Krippendorff's (2004) α metric. The error analysis suggests that the rich morphology and free phrase order nature of Bengali restricts the baseline system in capturing the *holder* and *topic* as well as disambiguating them in complex, compound, and passive sentences.

Thus, a Support Vector Machine (SVM) (Joachims, 1998) based supervised classifier is employed as well for co-reference identification. In this classifier, each of the input vectors containing *emotional expression*, associated *holder*, and *topic* is prepared from each of the annotated Bengali blog sentences. The feature vector is prepared based on the information present in the sentences containing lexical, syntactic, semantic, rhetorical, and overlapping features (word, part-of speech (POS), and Named Entity (NE)). Considering each of the input vectors as a unit to be coded in terms of the values of a variable, the standard Krippendorff's (2004) α metric produces a satisfactory score that outperforms the baseline system on the test set. This observation suggests that the adoption of error handling features, along with the features for syntax, semantics, and rhetorical structure, improves the performance of the co-reference identification reasonably. Different types of error cases have been analyzed, and we employed different rule-based post-processing techniques to solve the error cases. The rest of this paper is organized as follows. Section 2 describes the related work. The baseline system is described in Section 3. The supervised framework with feature analysis is discussed in Section 4. Experiments and associated results are specified in Section 5. The error analysis and post processing techniques are discussed in Section 6. Finally, Section 7 concludes the paper.

## 2. Related Work

The current trend in the emotion analysis area is exploring machine learning techniques (Sebastiani, 2002; Mishne & Rijke, 2006) that consider the problem as text categorization or analogous to topic classification, which underscores the difference between machine learning methods and human-produced baseline models (Alm *et al.*, 2005). The affective text shared task on news headlines for emotion and valence level identification at SemEval 2007 has drawn focus to this field (Strapparava & Mihalcea, 2007). In order to estimate affects in text, the model proposed by Neviarouskaya *et al.* (2007) processes symbolic cues and employs natural language processing techniques.

Prior work in identification of opinion *holders* has sometimes identified only a single opinion per sentence (Bethard *et al.*, 2004) and sometimes several opinions (Choi, 2005). Identification of opinion *holders* for Question Answering with a supporting annotation task was attempted in Wiebe *et al.* (2005). Before that, another work on labeling the arguments of the verbs with their semantic roles using a novel frame matching technique was carried out in Swier and Stevenson (2004). Based on the traditional perspectives, another work discussed in Hu *et al.* (2006) uses an emotion knowledge base for extracting the emotion *holder*. The machine learning based classification task for “not *holder*,” “weak *holder*,” “medium *holder*,” or “strong *holder*” is described in Evans (2007). Kim and Hovy (2006) identified the opinion holder with the topic from media text using semantic role labeling. An anaphor resolution based opinion *holder* identification method exploiting lexical and syntactic information from online news documents was attempted by Kim *et al.* (2007). The syntactic models of identifying the emotion *holder* for English emotional verbs were developed in Das and Bandyopadhyay (2010d).

In the related field of opinion *topic* extraction, different researchers have contributed their efforts (Kobayashi *et al.*, 2004; Nasukawa *et al.*, 2003; Popescu & Etzioni, 2005). Nevertheless, these works are based on lexicon look up and are applied to the domain of product reviews. The *topic* annotation task on the MPQA corpus is described in Stoyanov and Cardie (2008).

The method of identifying an opinion with its *holder* and topic from online news is described in Kim and Hovy (2006). The model extracts opinion *topics* associated with a specific argument position for subjective expressions signaled by verbs and adjectives. Similarly, the verb based argument extraction and associated *topic* identification is considered in the present system. Nevertheless, opinion topic identification differs from topic segmentation (Choi, 2000). The opinion *topics* are not necessarily spatially coherent as there may be two opinions in the same sentence on different *topics*, as well as opinions on the same *topic* that are separated by opinions that do not share that *topic* (Stoyanov & Cardie, 2008). The authors established such a hypothesis by applying the technique of co-reference identification for topic annotation. In the case of our present system, the building of fine-grained *topic* knowledge based on the rhetorical structure and segmentation of *topics* using different types of lexical, syntactic, and overlapping features substantially reduces the problem of emotion *topic* distinction. It must be mentioned that the proposed method obtains a moderately more reliable alpha score in comparison to some related results in Stoyanov and Cardie (2008a).

Moreover, all of the aforementioned works have been attempted for English. Recent study shows that non-native English speakers support the growing use of the Internet[1]. In addition to that, a rapidly growing number of web users from multilingual communities have focused attention on improving multilingual search engines in respect to sentiment or emotion. This raises the demand for emotion analysis for languages other than English. Bengali is the sixth most popular language in the World[2], second in India, and the national language in Bangladesh, but it is less computerized than English. Works on emotion analysis in Bengali have started recently (Das & Bandyopadhyay, 2009a; 2010a). The comparative evaluation of the features on equivalent domains for Bengali and English language can be found in Das and Bandyopadhyay (2009b). To the best of our knowledge, at present, no such user-topic co-reference analysis of emotion has been conducted for Bengali or for other Indian languages. Thus, we believe that this work would meet the demands of user-*topic* focused emotion analysis systems.

## 3. Baseline System

A simple rule-based system has been designed to identify the *emotional expression*, *holder*, and *topic* from the sentences and their co-references. A simple neighboring chunk consideration approach that assumes that *emotional expression*, *holder*, and *topic* appear as neighboring chunks in a sentence has been introduced to identify the co-reference among the three components. The details of the system are as follows.

***Identifying Emotional Expression*:** The blog sentences are passed through an open sourced Bengali shallow parser[3]. This shallow parser gives different morphological

[1] http://www.internetworldstats.com/stats.htm

[2] http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

[3] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

information (*root, lexical category of the root, gender, number, person, case, vibhakti, tam, suffixes, etc*.) that help in identifying the lexical patterns of the *emotional expressions*. The shallow parsed sentences are preprocessed to generate simplified lexical patterns (as shown below). We search through each of the component words from the chunks in the Bengali *WordNet Affect* lists (Das & Bandyopadhyay, 2010b). If any word present in a chunk is an emotion word (*e.g*. কৌতুক *koutuk* 'comic'), all of the words present in that extracted chunk are treated as the candidate seeds for an *emotional expression*. Identification of an *emotional expression* containing a single emotion word is straightforward. Nevertheless, we include all of the words of a chunk in order to identify long *emotional expressions*. Consecutive words that appear in a chunk and contain at least one emotion word also form an *emotional expression*. An example of a shallow parsed result follows.

((JJP **সুন্দর** **'sundar' [beautiful]** **JJ** <fs af='সুন্দর ,adj,,,,d,শুন্য,শুন্য'> )
( NP **কৌতুকটা** 'koutukta' [comic] **NN** <fs af=' কৌতুক ,v,,,,, টা , টা > ))

In many cases, the components of a given *emotional expression* are separated by stop words (*e.g*. একটি *ekti* 'a,' ঐ *oi* 'that,' এই *ei* 'this'), conjunctions (*e.g*. এবং *ebong* 'and,' অথবা *athoba* 'or,' কিন্তু *kintu* 'but,' *etc*.), negations (*e.g*. নয় *noy* 'not,' না *na* 'neither,' *etc*.) or intensifiers (তাই *tai* 'so,' খুব *khub* 'very,' কম *kam* 'less,' বেশি *beshi* 'much'). Each *emotional expression* is tagged with any of Ekman's (1993) six emotions, based on the type of the component word of the *emotional expression* present in the Bengali *WordNet Affect* lists.

***Identification of Emotion Holder:*** The baseline system considers the phrasal patterns containing similarity clues to identify the emotion *holders*. The patterns are grouped according to part-of-speech (POS) categories. It has been observed that the hints of grouping the patterns are present mostly in the user comment portions of the Bengali blog texts. The Bengali blog structure[4] has been designed well, and each of the user comment portions starts with a corresponding username. The username is the default hint that helps in capturing the first *holder* present in an anchoring vector representing the nested sources. In other cases, the POS tags of the shallow parsed sentences contain similar patterns at the lexical level. The Named Entities (NEs) that are tagged with NNPC (Compound proper noun), NNP (Proper noun), NNC (Compound common noun), NN (Common noun), or PRP (Pronoun) tags at the beginning of a sentence are tagged as the possible candidates for the emotion *holder*. The similarity pattern consists of two phrasal constituents, the subject and the verb. The portion of the sentence excluding the subject and the verb that contains the additional constituents of the

[4] www.amarblog.com/

sentence has been identified as the common portion (*Common_Portion*). As Bengali is a free phrase order language, the order between the verb and the common portion is not fixed.

A general POS level pattern, such as [<NNP/NNPC/NN/NNC/PRP> {<VBZ/VM><*Common_Portion*>}], is considered for capturing clues about an emotion *holder*. The components of the *Common_Portion* are assembled after the first occurring POS tags of types NNP, NNC, or PRP in the POS tagged sentence until the verb POS, like VBZ or VM, is reached. The remaining components present in the sentence after the verb are appended to the common portion (*Common_Portion*).

The similarity patterns mostly exist in simple sentences. Such information is difficult to obtain from complex or compound sentences under this scheme; thus, the system also fails to identify nested emotion *holder*s. A total of 59 complex sentences and 34 compound sentences were present in the test set of 503 sentences.

***Identifying Emotion Topic:*** The shallow chunked texts formed by removing *emotional expressions* and *holders* were identified as the responsible spans that contain one or more potential emotion *topics*. Without attempting any typical strategy, the words containing only the POS tags of NNP or NNC of the shallow chunks were identified as the emotion *topics*.

***Emotional Expression Holder / Topic Co-reference:*** The emotion *topic* is intended by the emotion *holder*, and the *topic* depends on the context in which its associated emotional expression occurs (Stoyanov & Cardie, 2008a). Based on this hypothesis, each identified *holder* and *topic* that is associated with an *emotional expression* in a sentence is termed as co-referent if it shares the same *emotional expression* with the others. A rule-based technique has been adopted to identify the co-reference between the *holder* and *topic* with respect to an *emotional expression* if the chunks that are responsible for emotion *holder* or *topic* are the immediately neighboring chunks of that *emotional expression*.

***Evaluation:*** The three identified components were stored in a vector. The evaluation of the vectors was carried out using Krippendorff's (2004) α metric by considering each of the vectors as the unit to be analyzed. Two vectors were filled up by the emotional components that were acquired from the annotated sentences and system generated results. Considering the annotated and system generated outputs as two separate raters, we used the number of identified components as the values to be assigned for each vector. We evaluated the system through the help of inter-rater agreement and measured the performance of the system using Krippendorff's (2004) α metric. The inter-rater agreement produces an α score of 0.53 on the test set. This is a standard metric employed for inter-annotator reliability studies. Krippendorff's α is a theoretically founded measure with a nice probabilistic interpretation. It was designed to measure the reliability of coding agreement, and the generalization of this metric was used as the evaluation metric for identifying co-reference in opinion *topic* annotation (Stoyanov & Cardie, 2008a). Krippendorff's alpha is applicable to any number of coders (each assigning one value to one unit of analysis); to incomplete (missing) data; to any number of values available for coding a variable; and to binary, nominal, ordinal, interval, ratio, polar, and circular metrics (Levels of Measurement). In addition, it adjusts itself to small sample sizes of reliability data. We have concentrated only on nominal alpha as we have considered the strings of names.

It is observed that some sentences may or may not contain all three emotional components. Hence, three out of four values for each of the raters are assigned based on the number of annotated or acquired emotional components from the gold standard and system-generated data, respectively. One value has been considered for undetermined cases. If no annotated or system generated emotional component is tagged or acquired, the corresponding vector unit is considered as incomplete or as containing missing data. The metric, nominal alpha, produces an α score of 0.53 for measuring the agreement between the annotated and system generated data. The lower score of α, along with the availability of different features in the corpus, also motivated us to adopt a machine learning framework.

## 4. Supervised Framework

The *Topic* co-reference resolution resembles another well-known problem in NLP - the noun phrase (NP) co-reference resolution that considers machine learning frameworks (Soon *et al.*, 2001; Ng & Cardie, 2002). Therefore, we adopted a Support Vector Machine (SVM) (Joachims, 1998) based standard machine learning approach for identifying *holder-topic* co-reference from the perspective of *emotional expressions*, where the input vectors contain emotional expressions, *holders*, and *topics*. The training and classification processes for SVM were carried out by YamCha toolkit[5] and TinySVM-0.07[6], respectively. The system was trained with 2234 sentences. The best feature set was identified using 630 development sentences. An Information Gain Based Pruning (IGBP) was applied to the development set, and it improved the performance of the supervised system significantly.

Feature plays a crucial rule in the SVM framework. By manually reviewing the blog data and different language specific characteristics, word level and context level features have been selected heuristically for our classification task. The heart of our method is to give an input vector containing the *emotional expression*, *holder*, and *topic*, and the goal of the classifier is to determine whether the co-reference exists among the available components of the vector or not. Therefore, we have considered five different classes for identifying the co-reference between any pair of components. The classes are Expression-*Holder* (EH), Expression-*Topic* (ET),

[5] http://chasen-org/~taku/software/yamcha/

[6] http://chasen.org/~taku/software/TinySVM/

Holder-*Topic* (HT), Expression-*Holder-Topic* (EHT), and *none*. We use the manually annotated corpus (Das & Bandyopadhyay, 2010a) to train the classifier automatically. We construct each training example for each input vector. The co-reference identification relies on the expressiveness of the features used to describe the training example. We use the following four categories of features: lexical, syntactic, semantic, rhetorical, and overlapping features.

## 4.1 Lexical Features

***Parts-of-Speech*** **(POS)**: We are interested in the *noun*, *adjective*, *verb*, and *adverb* words as these are emotion informative constituents. The POS features are extracted from the shallow parsed results used by the baseline system.

***Negations*** **(NEG)**: Negative words that are annotated in the corpus (Das & Bandyopadhyay, 2010a) (নয় *noy* 'not,' না *na* 'neither,' *etc*.) were considered as a separate feature.

***Conjunctions*** **(CONJ)**: The *Conjunctions* were annotated in the emotion corpus (Das & Bandyopadhyay, 2010a). The conjunctions were used as features (*e.g*. এবং *ebong* 'and,' অথবা *athoba* 'or,' কিন্তু *kintu* 'but,' *etc*.) for training and testing.

***Punctuation Symbols*** **(Sym)**: Symbols, such as (,), (!), (?), are often used in single or multiple numbers to emphasize *emotional expressions* and are considered crucial clues for identifying emotional presence in a sentence. Thus, a special feature for such symbols was added in the active feature set for training and testing of the supervised system.

***Emoticons*** **(emot_icon)**: Emoticons (☺,☹,😐) and their consecutive occurrence generally contribute real sentiment to the *emotional expressions* that precede or follow them. Like punctuation symbols, emoticons were also included as a feature. A knowledge base for emoticons has been prepared by experts after minutely analyzing the Bengali blog data. Each image link of an emoticon in the raw corpus was mapped to its corresponding textual entity in the tagged corpus with its proper emotion types using the knowledge base (Das & Bandyopadhyay, 2009).

## 4.2 Extraction of Subcategorization Frames for Identifying Syntactic Features

We augment the knowledge of subcategorization frames or syntactic frames for identifying emotion *holders* and *topics*. The identification of syntactic frames is not straightforward. The detailed methodology is as follows.

***Verb Identification:*** The words tagged as main verb (VM) and belonging to the verb group chunk (VGNF) in the corpus are identified (*e.g.* ভালোবাসা *bhalobasa* 'love') as simple verbs from the shallow parsed sentences. In cases of compound or conjunct verbs, patterns like {[XXX] (NN) [YYY] (VM)} are retrieved (*e.g*. VGNF {[আনন্দ *ananda*] (NN) [করা *kara*] (VM)} means *enjoy*). The light verbs [YYY] tagged with 'VM' generally occur in any inflected form. Different suffixes may be attached to a simple verb or light verb depending on various features, like tense, aspect, and person. An in-house Bengali stemmer with an accuracy of 97.09% used a suffix list to identify the stem forms of the simple and light verbs.

***English Equivalent Synset Identification:*** The determination of an equivalent English synset of a Bengali verb was carried out using a Bengali to English bilingual dictionary[7]. The method to extract the English equivalent synsets of the Bengali verbs was based on the work done in Banerjee *et al.* (2009). We have identified the English equivalent verb synsets of the Bengali verb entries that are present in the dictionary. For example, the dictionary entries for the conjunct verb আনন্দ করা *ananda kara* 'enjoy' are as follows.

< আনন্দ করা **v**. to ***rejoice***; to ***make merry***….>

Different synonyms for a Bengali verb having the same sense are separated using "," and different senses are separated using ";" in the dictionary. The synonyms, including similar senses of the target verb, were extracted from the dictionary and yielded a set called the English Equivalent Synset (EES). In the above example, two English Equivalent Synsets (EES) are extracted for the conjunct verb আনন্দ *ananda* করা *kara* 'enjoy'.

***English Equivalent Frame Identification:*** It also has been found that each of the English Equivalent Synsets (EES) occurs in each separate class of English VerbNet (Kipper-Schuler, 2005). VerbNet associates the semantics of a verb with its syntactic frames and combines traditional lexical semantic information, such as *thematic roles* and *semantic predicates*, with *selectional restrictions*. Member verbs in the same VerbNet class share common syntactic frames; thus, they are believed to have the same syntactic behavior. The VerbNet files containing member verbs and possible subcategorization frames are stored in XML file format. Hence, the XML files of VerbNet were pre-processed to build up a general list that contains all verbs, their classes, and possible subcategorization frames (primary as well as secondary). This preprocessed list was searched to extract the present subcategorization frames for each verb (*e.g. love*) of the English Equivalent Synsets (EES) (*e.g.* love) corresponding to the Bengali verb. These extracted subcategorization frames are believed to be the valid set of argument structures for the Bengali verbs (Banerjee *et al.*, 2010).

***Frame Matching:*** On the other hand, the shallow parsed Bengali sentences are passed through a rule based *phrasal-head* extraction module to identify the phrase level argument structures of the sentences corresponding to the position of the verbs. The extracted *head part* of every

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

phrase from a parsed sentence is considered as a component of its sentential argument structure. If an acquired argument structure for a Bengali emotional sentence is matched with any of the available extracted frames of English VerbNet, the *thematic role* based *holder* (*Experiencer, Agent, Actor, Beneficiary*, *etc.*) and *topic* (*Topic*, *Theme*, *Event, etc.*) information associated with the English frame syntax is mapped to the appropriate slot of the acquired Bengali argument structure. Tag conversion routines were developed to transform the POS of the system generated argument structures into the POS of the VerbNet frames. The phrase level similarity between these two languages helps in identifying the subcategorization frames (Banerjee *et al.*, 2009). An example follows:

*রাশেদ* অনুভব করেছিল যে রামের সুখ অন্তহীন
(*Rashed*) (*anubhob*) (*korechilo*) (*je*) (*Ramer*) (*sukh*) (*antohin*)
*Rashed felt that Ram's pleasure is endless.*
*Vector*: < EH_রাশেদ, EH_*রাম*, ET_সুখ >
Acquired Argument Structure: [NNP VM DET-*je* S]

The argument structure contains a sentential complement "S" started by যে –*je* with DET type POS. The argument structure is acquired for the Bengali conjunct verb অনুভব করা *anubhab kara* 'feel'. One of the extracted VerbNet frame syntax containing –*that* type sentential complement for the equivalent English verb *feel* is as [<NP value="*Experiencer*" > </VERB> < S-*that* (Sentential –*that* Complement)>]. As the acquired argument structure matches the extracted VerbNet frame syntax, the *holder* related roles (*e.g. Experiencer*) associated with the VerbNet frame was mapped to the equivalent phrase in the acquired argument structure of the Bengali sentence. The phrase (রাশেদ) is now considered as a candidate of emotion *holder*. Additionally, the sentential complement portion is also passed through the syntactic model for obtaining any implicit emotion *holders*. The case markers in Bengali are required to identify the emotion *holder*s as the case markers give the useful hints to capture the *selectional restrictions* that play a key role in distinguishing the emotion *holder*s from other valid alternatives.

## 4.3 Semantic Features

***Emotion/Affect Words* (EW)**: The presence of a word in the Bengali *WordNet Affect* lists (Das & Bandyopadhyay, 2010b) identifies the emotion/affect words. The tagged affect words are considered as both lexical and semantic features in the case of handling the *emotional expressions*.

***Intensifiers* (INTF)**: The Bengali *SentiWordNet* was developed by replacing each word entry in the synonymous set of the English *SentiWordNet* (Esuli & Sebastiani, 2006) by its possible

Bengali synsets using the English to Bengali bilingual dictionary that was developed as part of the EILMT project[8]. The chunks containing JJ (adjective) and RB (adverb) tagged elements were considered to be intensifiers. If the intensifier was found in the *SentiWordNet*, then the positive and negative scores of the intensifier were retrieved from the *SentiWordNet*. The intensifier is classified into the list of positive (pos) (**INTF*pos***) or negative (neg) (**INTF*neg***), for which the average retrieved score is higher. The intensifiers play an important role in identifying the lexical association among the component words of an *emotional expression* and linking the emotion components based on their POS.

***Multiword Expressions:*** *Reduplicated* words (সন্দ সন্দ *sanda sanda* [doubt with fear]) and *Idioms* (তাসের ঘর *taser ghar* [weakly built], গৃহদাহ *grrihadaho* [family disturbance]), which were annotated in the Bengali emotion blog corpus (Das & Bandyopadhyay, 2010a), have been considered as semantic features for the *emotional expressions*.

## 4.4 Rhetoric Features

The present task acquires the rhetorical components, such as *locus*, *nucleus*, and *satellite* (Mann & Thompson, 1988), from a sentence, as these rhetorical clues help in identifying the individual *topic* spans. The part of the text span containing an annotated *emotional expression* is considered as *locus*. Primarily, the separation of *nucleus* from *satellite* is done based on the punctuation marks (,), (!), (?). Frequently used *discourse markers* (যেহেতু *jehetu* 'as,' যেমন *jemon* '*e.g.*,' কারণ *karon* 'because,' মানে *mane* 'means' ) and *causal verbs* (ঘটায় *ghotay* 'caused') also act as useful clues if they are explicitly specified in the text. Stoyanov and Cardie (2008a) mentioned that the *topic* depends on the context in which its associated *emotional expression* occurs. If any word of an *emotional expression* co-occurs with any word element of the *nucleus* or *satellite* in the same shallow chunk, the feature is considered a *common rhetoric similarity*. Otherwise, the feature is considered a *distinctive rhetoric similarity*. The features aim to separate emotion *topics* from non-emotion *topics* as well as the individual *topic* from an overlapped *topic* region.

## 4.5 Overlapping Features

***Word Overlap:*** This feature is *true* if any two *topic* spans contain a common word.

**Part-of-Speech Overlap:** The verb, noun, adjective, and adverb are considered as overlapping informative constituents.

***NP Co-reference:*** This binary feature is *True* if the two chunks contain NPs that are determined to be co-referent by applying a rule of *common rhetoric similarity*.

[8] 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.

***Named Entity (NE)***: Each of the sentences is passed through a Named Entity Recognizer (Ekbal & Bandyopadhyay, 2008) to identify the named entities in that sentence.

If any word is tagged as a named entity (NE), a feature is assigned for either emotion *holder* or *topic*. If, however, the word is present in *satellite* and not tagged as an emotion *holder* (*EH*) feature, the word is selected as a potential candidate for *topic*. This distinguishing feature is considered for identifying the *holder* and *topic* separately from an NE overlapped context.

## 5. Experimental Results

The combination of multiple features in comparison with a single feature generally shows a reasonable performance enhancement of any classification system. The impact of different features and their combinations was measured on the development set of 630 sentences. Different unigram and bi-gram context features (word and POS tag level) and their combinations were generated from the training corpus as well. We added each feature into the active feature list one at a time to see if the inclusion of a feature in the existing feature set improved the *F-Score* of the system on the development set. The final active feature set was applied to the test data. During the SVM-based training phase, the current token word with the three previous and three following words and their corresponding POS, along with negation or intensifier, were selected as context features for that word. We used Krippendorff's (2004) alpha (as discussed in Section 3) for measuring the performance of the system. The importance of incorporating the features was examined through Information Gain (*InfoGain*). All of the results were obtained by the 10 fold cross validation method.

***Information Gain Based Pruning (IGBP):*** This decision technique was used to measure the importance of a feature (X) with respect to the class attribute (Y). Formally, the information gain of a feature X with respect to a class attribute Y is the reduction in uncertainty about the value of Y when we know the value of X:

$$\mathit{InfoGain}\,(Y;X) = \mathit{entropy}(Y) - \mathit{entropy}(Y|X)$$

where X and Y are discrete variables taking values $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$, respectively. The *Entropy*(Y) is defined as:

$$\mathit{Entropy}\,(Y) = - \sum_{i=1 \text{ to } n} P(Y=_{yi})\, \mathit{log}_2\, P(Y=_{yi})$$

The conditional entropy of Y given X is defined as:

$$\mathit{Entropy}\,(Y|X) = - \sum_{i=1 \text{ to } m} P(X=_{xj})\, \mathit{Entropy}\,(Y|X=_{xj})$$

The features with high Information Gain (*InfoGain*) reduce the uncertainty about a class to the minimum. In our experiment on the development set, all of the features except the features for the *causal verbs* and *distinctive rhetoric similarity* achieved a high Information Gain (*InfoGain*). The word features (*e.g.* non-emotional words, such as *gather*, *seem*, *etc.*) were not considered based on a threshold of 0.5.

The metric, nominal alpha produced an $\alpha$ score of 0.53 between the annotated and system generated data. Generally, the alpha $\alpha$ score aims to probabilistically capture the agreement of annotated data and separate it from the chance of agreement. The baseline score achieved for the overall agreement was 0.53, which is below the generally accepted level, while $\alpha$ for the supervised system was 0.63, which is moderately acceptable and reliable. The scores of $\alpha$ for the baseline system and supervised system, along with some important features, their combinations, and pruning steps, are shown in Table 1. The $\alpha$ score loses its probabilistic interpretation due to the way it is adapted to the problem of co-reference classification. It is observed that the score of $\alpha$ increased rapidly while considering the syntactic, rhetorical, and overlapping features. The overlapping features also cause problems because of the free phrase order characteristics of the Bengali language. The overlapping context of *emotional expression* and *topic* generates errors. Nevertheless, the application of Named Entities (NEs) reduces the problem of distinguishing *holder* and *topic*.

***Table 1. Krippendorfff's $\alpha$ for different feature combinations and pruning.***

| Features | Krippendorff's $\alpha$ |
|---|---|
| **Baseline System** | **0.5344** |
| Supervised System (Lexical Features) | 0.3561 |
| Supervised System (Syntactic Features) | 0.4002 |
| Supervised System (Semantic Features) | 0.3215 |
| Supervised System (Rhetorical Features) | 0.4176 |
| Supervised System (Overlapping Features) | 0.2345 |
| Supervised System (Lexical+Syntactic) | 0.4890 |
| Supervised System (Syntactic+Rhetoric) | 0.5012 |
| Supervised System (Syntactic+Semantic+Rhetoric) | 0.5201 |
| Supervised System (Lexical+Syntactic+Rhetoric) | 0.5421 |
| **Supervised System (Lexical+Syntactic+Semantic+Rhetoric+Overlapping)** | **0.6121** |
| **Supervised System (All Features) + IGBP** | **0.6332** |

## 6. Error Analysis

The error analysis was conducted on the development set of 630 sentences. We incorporated different rule-based post processing techniques for handling the error cases, and the system achieved an alpha score of 0.67. Four types of error cases were identified, and four different rules were proposed to reduce the error cases.

***Case 1:*** *Appositive Use:* The implicit emotion *holders* may be present in a sentence. (*e.g.* রাম 'Ram' in the case of রামের সুখ 'Ram's pleasure'). The identification of the emotion *holder* at the sentence level requires the knowledge of two basic constraints (*implicit* and *explicit*) separately. The *explicit* constraints identify the single prominent emotion *holder* that is directly involved with the *emotional expression*, whereas the *implicit* constraints identify all direct and indirect nested sources as emotion *holder*s. The following example contains the emotion *holder* নাসরিন সুলতানা (*Nasreen Sultana*) based on *implicit* constraints.

*Holder:* < গেদু চাচা, নাসরিন সুলতানা >
**গেদু চাচা** বলে, না গো বোন , আমি **নাসরিন সুলতানার**
(**Gedu ChaCha**) (bole) : (na) (go) (bon) , (ami) (**Nasreen Sultanar**)
দুঃখের কথাতে *কেঁদে* ফেলি।
(dookher) (kathate) (*kende*) (feli)
***Gedu Chacha** says, no my sister, I fall into cry on the sad speech of **Nasreen Sultana**.*

Solution: We considered the suffixes that are determined from the shallow parsed phrases to identify the appositive cases. In the above example, the appositive case (*e.g.* রামের সুখ (*Ram's pleasure*)) is also identified and placed in the vector by removing the inflectional suffix (-এর *-er* in this case). Sometimes, the vibhakti and tam information also play effective roles in identifying emotion holders.

***Case 2:*** *Anaphoric Presence of Holders*: Another similar problem is identified in the above example. The emotion *holders* are sometimes referred to via anaphors. Sometimes, the candidate anaphors are linked with the *emotional expressions* instead of the actual emotion *holders*. The actual emotion *holder* গেদু চাচা 'Gedu ChaCha' expresses the emotion in a clause that is represented by the anaphor আমি ami 'I' in another clause.

Solution: The sentences of user comments in the adopted blog corpus contain a special default phrasal pattern that helps in identifying the emotion holders ([<Named Entity> <say>] *e.g.* গেদু চাচা বলে: (*Gedu ChaCha bole*), রাশেদ বলেছেন (*Rashed bolechen*)**,** and সায়ন বলেছে (*Sayan bolechhe*)). Hence, if a pronoun is present with an *emotional expression*, the preceding Named Entities of such a default phrasal pattern are considered as the emotion *holders*.

***Case 3:*** *Multiple Holders and Topics:* The complex or compound sentences contain more than one clause, and each of the clauses may contain individual *emotional expressions*. The *holders* and *topics* associated with the *emotional expressions* in all of the clauses need fine-grained study of the sentential structures. The following example shows that two *emotional expressions* (দুঃখ *dookkha* 'sorrow' and আনন্দ *ananda* 'happy') contain two different *holders* (গেদু চাচা *Gedu ChaCha* and চাচি *Chachi*).

**গেদু চাচার** ***দুঃখ*** থাকা সত্ত্বেও **চাচি** ***আনন্দ*** করে সবাইকে
(**Gedu ChaChar**) (***dookkha***) (thaka) (satweo) (**Chachi**) (***ananda***) (kare) (sabaike)
নিয়ে থাকে ।
(niye) (thake)
*Though **Gedu Chacha** has sorrow, **Chachi** lives happily with all.*

Solution: As the complex or compound sentences contain more than one clause and each of the clauses contains individual *emotional expressions*, we consider the sentential rhetorical structure. Instead of identifying rhetorical relations (Mann & Thompson, 1988), the present task acquires the rhetorical components, such as *locus*, *nucleus*, and *satellite* from a sentence, as these rhetoric clues help in identifying the individual *holder* and *topics* associated in each clause of the sentence. The part of the text span containing the *emotional expression* is considered as *locus*. Primarily, the separation of *nucleus* from *satellite* is done based on the punctuation marks (,), (!), (?). Frequently used *discourse markers* (যেহেতু *jehetu* 'as,' যেমন *jemon* '*e.g.*,' কারণ *karon* 'because,' মানে *mane* 'means') and *causal verbs* (ঘটায় *ghotay* 'caused') also are useful clues if they are explicitly specified in the text and present in a manually prepared seed list. If any word in the *emotional expression* co-occurs with any word element of the *nucleus* or *satellite* in the same chunk, the feature is considered a *common rhetoric similarity*. Otherwise, the feature is considered a *distinctive rhetoric similarity*. The chunks identified by the syntactic system as the holder and topic and tagged as *common rhetoric similarity* are only considered for each of the clauses of a sentence. For this reason, all possible holders and topics associated to all of the clauses of a sentence are identified by the syntactic system.

***Case 4:** Overlapping Topic Spans:* It is observed that the emotion *topics* containing single word tokens are identified more easily than multi word *topics*. Sometimes, the emotion related *topics* coexist with other potential non-emotional *topics*. As the *topics* may consist of multi-word strings, the text spans denoting the *topic* spans create problems in identifying emotion *topic* span from other non-emotional *topic* spans. In the following example, the *emotional expression* আনন্দ *ananda* 'enjoy' is related to the topic গান *gan* 'song' and টিভি *TV* 'television'. The baseline system additionally captures বই *boi* 'book' that is a potential but non-emotion *topic*.

তুমি তো বই পড়তেই না, এখন দেখছি তুমি **গান**, **টিভি** তেও

(tumi) (to) (boi) (portei) (na), (ekhon) (dekhchi) (tumi) (**gan**), (**TV**) (teo)

*আনন্দ* পাওনা।

(*ananda*) (paona)

*You never used to read books; now we notice that you also don't enjoy* ***song/ television*****.**

Solution: The *topic* of an opinion depends on the context in which its associated *opinion expression* occurs (Stoyanov & Cardie, 2008a). The *common rhetoric similarity* feature helps the syntactic system by aiming to separate emotion *topics* from non-emotion *topics* and to separate the overlapping possibilities of discrete emotion topic spans from non-topical contiguous regions. If the identified *topic* chunks are tagged with *common rhetoric similarity*, the chunks are classified as emotional *topics* and separated from non-topical elements in a sentence. The improvements at some important steps by incorporating the rule based post-processing techniques are shown in Table 2. It is observed that the simple rules have substantially reduced errors and have improved the performance of the system satisfactorily. The application of the post-processing techniques also achieves an alpha score of 0.6721 on the test set.

***Table 2. The alpha scores of the system after handling the four error cases.***

| Cases | Krippendorff's $\alpha$ |
|---|---|
| Before Error Analysis | 0.6332 |
| Case 1 | 0.6476 |
| Case 2 | 0.6417 |
| Case 3 | 0.6498 |
| Case 4 | 0.6402 |
| Case 1+ Case 2 | 0.6510 |
| Case 1+Case 3 | 0.6533 |
| Case 1+Case 2+Case 3 | 0.6601 |
| Case 1+Case 2+Case 4 | 0.6625 |
| Case 1+Case 3+Case 4 | 0.6678 |
| Case 1+Case 2+Case 3+Case 4 | **0.6772** |

## 7. Conclusion

The automatic extraction of *emotional expressions*, sentential emotion *holders*, and *topics* from Bengali blog data is accomplished in the present task. The supervised implementation of the system shows improvement over the rule-based baseline because the rule-based system fails to capture the implicit textual clues whereas the supervised system captures the clues in terms of combined features. The evaluation of the co-reference using Krippendorff's alpha is helpful in diagnosing the importance of the three emotional components. The rule-based post-processing techniques for reducing the error cases have shown substantial improvement in the performance of the system. From the overall analysis, it is observed that the identification of emotional co-reference is helpful in identifying user-topic relations. The handling of metaphors and their impact in detecting sentence level emotion is not considered. Future analysis concerning the time based emotional change can be used for *topic* model representation. The need for co-reference requires that the presence of indirect affective clues can also be traced with the help of the *holder* and *topic*.

## Reference

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. *HLT- EMNLP*, 579-586.

Banerjee, S., Das, D., & Bandyopadhyay, S. (2009). Bengali Verb Subcategorization Frame Acquisition - A Baseline Model. *ACL-IJCNLP-2009*, *ALR-7 Workshop,* 76-83.

Banerjee, S., Das, D., & Bandyopadhyay, S. (2010). Classification of Verbs – Towards Developing a Bengali Verb Subcategorization Lexicon. *GWC*, 76-83.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., & Jurafsky, D. (2004). Automatic Extraction of Opinion Propositions and their Holders, In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of HLT/EMNLP*.

Das, D., & Bandyopadhyay, S. (2009a). Word to Sentence Level Emotion Tagging for Bengali Blogs. *ACL-IJCNLP 2009,* 149-152.

Das, D., & Bandyopadhyay, S. (2009b). Emotion Tagging – A Comparative Study on Bengali and English Blogs. *ICON-09*. 177-184.

Das, D., & Bandyopadhyay, S. (2010a). Labeling Emotion in Bengali Blog Corpus – A Fine Grained Tagging at Sentence Level. *ALR8, COLING-2010*, 47-55.

Das, D., & Bandyopadhyay, S. (2010b). Developing Bengali WordNet Affect for Analyzing Emotion. *ICCPOL-2010*, 35-40.

Das, D., & Bandyopadhyay, S. (2010c). Sentence Level Emotion Tagging on Blog and News Corpora. *Journal of Intelligent System (JIS),* 19 (2), 125-134.

Das, D., & Bandyopadhyay, S. (2010d). Emotion Holder for Emotional Verbs – The role of Subject and Syntax. In *CICLing*, A. Gelbukh (Ed.), LNCS 6008, 385-393.

Ekbal, A., & Bandyopadhyay, S. (2008). Named Entity Recognition using Appropriate Unlabeled Data, Post-processing and Voting. In *Informatica Journal of Computing and Informatics*, ACTA Press.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48(4), 384-392.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, *Language Resource and Evaluation Campaign*.

Evans, D. K. (2007). A low-resources approach to Opinion Analysis: Machine Learning and Simple Approaches, *NTCIR*.

Hu, J., Guan, C., Wang, M., & Lin, F. (2006). Model of Emotional Holder. *In Shi, Z.-Z., Sadananda, R. (eds.) PRIMA 2006. LNCS (LNAI)*, 4088, 534-539.

Joachims, T. (1998). Text Categorization with Support Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning*, 137-142.

Kim, Y., Jung, Y., & Myaeng, S.-H. (2007). Identifying Opinion Holders in Opinion Text from Online Newspapers. In *2007 IEEE International Conference on Granular Computing*, 699-702, doi:10.1109/GrC.2007.45.

Kim, S. M., & Hovy, E. (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., & Fukushima, T. (2004). Collecting evaluative expressions for opinion extraction. *IJCNLP*.

Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania.

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. *Thousand Oaks, CA: Sage*.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, *TEXT* 8, 243-281.

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

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007). Narrowing the Social Gap among People Involved in Global Dialog: Automatic Emotion Detection in Blog Posts, *ICWSM*.

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to co-reference resolution. In *Proceedings of ACL*.

Popescu, A., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP*.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1-47.

Seki, Y. (2007). Opinion Holder Extraction from Author and Authority Viewpoints. In *Proceedings of the SIGIR'07*, *ACM 978-1-59593-597-7/07/0007*.

Soon, W., Ng, H., & Lim, D. (2001). A machine learning approach to co-reference resolution of noun phrases. *Computational Linguistics*, 27(4), 521-544.

Stoyanov, V., & Cardie, C. (2008a). Annotating topics of opinions. In *Proceedings of Language Resource and Evaluation Campaign*.

Stoyanov, V., & Cardie, C. (2008b). Topic Identification for Fine-Grained Opinion Analysis. *Coling 2008*, 817-824.

Strapparava, C., & Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text, *ACL*.

Swier, R. S., & Stevenson, S. (2004). Unsupervised Semantic Role Labelling. *EMNLP*.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2), 1-54.

Yang, C., Lin, K. H.-Y., & Chen, H.-H. (2007). Emotion classification Using Web Blog Corpora. In *Proceedings of the IEEE, WIC, ACM International Conference on Web Intelligence*, 275-278.

Yang, C., Lin, K. H.-Y., & Chen, H. H. (2009). Writer Meets Reader: Emotion Analysis of Social Media from both the Writer's and Reader's Perspectives. In *Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 287-290.

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of ICDM*.

Zhang, Y., Li, Z., Ren, F., & Kuroiwa, S. (2008). A preliminary research of Chinese emotion classification model. *IJCSNS*, 8(11), 127-132.