Classifying Emotion in News Sentences: When Machine Classification Meets Human Classification

Plaban Kumar Bhowmick, Anupam Basu and Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur
West Bengal, India – 721302
{plaban, anupambas, pabitra}@gmail.com

Abstract : Multiple emotions are often evoked in readers in response to text stimuli like news article. In this paper, we present a method for classifying news sentences into multiple emotion categories. The corpus consists of 1000 news sentences and the emotion tag considered was anger, disgust, fear, happiness, sadness and surprise. We performed different experiments to compare the machine classification with human classification of emotion. In both the cases, it has been observed that combining anger and disgust class results in better classification and removing surprise, which is a highly ambiguous class in human classification, improves the performance. Words present in the sentences and the polarity of the subject, object and verb were used as features. The classifier performs better with the word and polarity feature combination compared to feature set consisting only of words. The best performance has been achieved with the corpus where anger and disgust classes are combined and surprise class is removed. In this experiment, the average precision was computed to be 79.5% and the average class wise micro F1 is found to be 59.52%.

I. INTRODUCTION

Social computing is a new age research area inspired by the human-human interactions in society. Human behavior, reflected through their expressions, gesture or spoken and written language, is profoundly influenced by their interactions with the society from birth to death. Consequently, the behavior of a person largely depends on the behavior of the other persons in the society. Sometimes the decision making process considers cues from the social contexts. For example, one tends to inclined to a particular product provided that other social actors are affined to the same. The new age digital systems are trying to exploit this role of social context in a number of applications like recommender system, intelligent tutoring system and many others.

Human Centered Computation (HCC) is a very recent area of thrust as the human world is getting more and more digital and one wants the digital systems to behave as close as to a human being. This requires the devices or the information processing systems to model the human behavior precisely. Emotion or affect is one aspect of human behavior which plays an important role in human perceptions and decision making thus influencing the way people interact in the society. In human-computer interaction, the computer interfaces need to recognize the affect of the end users in order to exhibit a truly intelligent behavior.

Expression or change of behavior is the most visible and prominent clues for recognizing emotion. Facial expressions [1], speech expressions [2] have widely been used in detecting emotion. Emotion is not a linguistic entity [3]. However, language is one of the most common modes for expressing emotion whether it is day-to-day speech communications (spoken language) or published communications (written language).

Like other communication entities, emotion communication involves two types of social actors: *speaker/writer* who sends some communication signals and *hearer/reader* who tries to evaluate the same. Two types of emotional evaluation may be possible in this context:

Writer/Speaker perspective evaluation: Given a text segment, the evaluation process predicts the emotion that the writer/speaker has intended to express. For example, the following expression suggests that the speaker is in angry state.

'Get out of my sight! Never come back again.'

Reader/Hearer perspective evaluation: This process predicts the emotions evoked in the reader's or hearer's mind in response a stimulus (text segment). For example, the following text segment may evoke sad emotion in the hearer's mind.

'The boy was so shocked to see his father dying in front of his eyes.'

In previous works [4-9], few attempts towards writer perspective analysis of emotion in text data

have been made. In all these studies, it has generally been assumed that the writer expresses only one emotion for a text segment. However, evocation of a blend of emotions is common in reader in response to a stimulus. For example, the following may evoke fear and sad emotion in readers mind.

'Militant attack kills over 30 persons in Nigeria.'

In this work, we perform reader perspective emotion analysis in text data where one text segment may evoke more than one emotion in reader. News is a media where certain facts in the articles are presented to the readers with the expectation that the articles evoke some emotional responses in the readers. So, this media is one potential data source for the computational study of reader perspective emotion. The problem of emotion classification of text can be stated as follows.

Definition 1. Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of sentences and $E = \{e_1, e_2, \ldots, e_k\}$ be the set of emotion classes (e.g., happy, sad etc.). The task is to find a function $h : S \rightarrow 2^E$, where 2^E is the powerset of E.

The problem of reader emotion classification in text data can be mapped to a multi-label text categorization problem. Multi-label classification algorithms have been categorized into two classes: algorithm adaptation methods and problem transformation methods. In algorithm adaptation methods, existing single label classification algorithms are adapted to handle multi-label data whereas, the multi-label data instances are transformed into single label by some transformation techniques (see [10]).

II. RELATED WORKS

Classifying emotion from reader's perspective is a challenging task and research on this topic is relatively sparse as compared to writer perspective analysis.

Affective text analysis was the task set in SemEval-2007 Task 14 [12]. A corpus of news headlines extracted from Google news and CNN was provided. Two types of tasks were to classify headlines into positive/negative emotion category as well as distinct emotion categories like anger, disgust, fear, happiness, sadness and surprise.

UPAR7 [13] is a linguistic rule-based approach towards emotion classification. The system performs emotion analysis on news headline data provided in SemEval-2007 Task 14. In the preprocessing step, the common words are decapitalized with the help of parts of speech tagger and Wordnet [14]. Each word first is rated with respect to emotion classes. The main theme word is detected by parsing a headline and it is given a higher weight than the other words in the headline. The emotion score boosting to the nouns are performed based on their belongingness to some general categories in Wordnet. The word scoring also considers some other factors like human will, negation and modals, hightech names, celebrities etc. The average accuracy, precision and recall of the system are 89.43%, 27.56% and 5.69%.

The system UA-ZBSA [15] gathers statistics from three different search engines (MyWay, AllWeb and Yahoo) to attach emotion labels to the news headlines. The work computes the Point wise Mutual Information (PMI) score of each content word of a headline with respect to each emotion by querying the search engines with the headline and the emotion. The accuracy, precision and recall of the system is reported to be 85.72%, 17.83% and 11.27%.

The system SWAT [16] adopts a supervised approach towards emotion classification in news headlines. The system develops a word-emotion map by querying the Roget's New Millennium Thesaurus [17]. This map is used to score each word in the headline and the average score of the headline words are taken into account while labeling it with a particular emotion. The reported classification accuracy, precision and recall are 88.58%, 19.46% and 8.62%.

The work by Lin and Chen [18, 19] provides the method for ranking reader's emotions in Chinese news articles from Yahoo! Kimo News. Eight emotional classes are considered in this work. Support Vector Machine (SVM) has been used as the classifier. Chinese character bigram, Chinese words, news metadata, affix similarity and word emotion have been used as features. The best reported system accuracy is 76.88%.

III. MULTI-LABEL CLASSIFICATION

In the task of multi-label classification problem, one example or instance may belong to more than one category simultaneously. Below we provide a formal definition of the task.

Definition 2. Let $X = \{x_1, x_2, \ldots, x_m\}$ be the data set and $\gamma = \{y_1, y_2, \ldots, y_k\}$ be the set of labels. The objective of the multi-label classification task is to find a hypothesis $h : X \rightarrow 2^{\gamma}$ with minimum error.

It is evident from definition 1 and definition 2 that emotion classification task can be mapped to multi-label classification. A number of multi-label classification algorithms are available in the literature and they have been applied to various classification tasks like text categorization, music emotion classification and many others. A comprehensive study and comparisons of the multi-label classification methods can be found in [10]. Among these, ADTboost.MH [11] is a multi-label

classification algorithm that has been used efficiently in text classification.

3.1. Multi-Label Alternating Decision Tree Learning: ADTboost.MH

In this section, we provide a brief overview of ADTboost.MH [11]. It is an Alternating Decision Tree (ADT) (variant of decision tree) based algorithm and is derived from ADTboost [20] and AdaBoost.MH [21] algorithms for handling multi-label classification problem. ADTboost.MH learns a multi-label Alternating Decision Tree (multi-label ADTree) consisting of a set of rules (R). These rules are learned through a number of boosting rounds provided as parameter. Note that the number of rules is same as the number of boosting steps. A rule in an ADTree is defined by a precondition C_1, a condition C_2 and two vectors of real numbers $(a_l)_{l\in\gamma}$ and $(b_l)_{l\in\gamma}$:

if C_1 *then (if* C_2 *then* $(a_l)_{l\in\gamma}$ *else* $(b_l)_{l\in\gamma}$ *else* $(0)_{l\in\gamma}$

A multi-label ADTree maps each instance to a vector of real numbers in the following manner:

Definition 3. A rule r: *if* C_1 *then (if* C_2 *then* $(a_l)_{l\in\gamma}$ *else* $(b_l)_{l\in\gamma}$ *else* $(0)_{l\in\gamma}$ associates a real value *r(x,l)* with any $(x,l)\in X\times\gamma$. if *(x,l)* satisfies $C = C_1 \wedge C_2$ then *r(x, l) = al*; if *(x,l)* satisfies $C = C_1 \wedge \neg C_2$ then *r(x l) = bl*; otherwise *r(x, l) = 0*.

An ADTree $R = r_i,\ i \in I$ (I = number of boosting rounds) associates a prediction value R(x,l) = $R(x,l) = \sum_{i\in I} r_i(x,l)$ with any *(x,l)*. A multi-label classification hypothesis is associated with H defined by *H(x,l) = sign(R(x,l))* and the real number *|R(x,l)|* is interpreted as the confidence assigned to *H(x,l)*.

Details of rule learning and weight updating procedures for ADTboost.MH are provided in [11].

3.2. Evaluation Measures in Multi-Label Classification

The idea of multi-label classification is different from that of single-label classification. Consequently, the evaluation measures applicable to single label classification are not relevant to the evaluation of multi-label classification task. We evaluate our emotion classification task with respect to four evaluation measures: *Hamming Loss (HL)* [11], *One Error (OE)*, *Coverage (COV)* and *Average Precision (AvP)* [21]. In addition, we report average class wise micro average F1 (*micro-F1*) value as another performance measure.

Let T be the test data set containing examples (t_i, Y_i), $i = 1, 2, \ldots, |T|$, $Y_i \subseteq \gamma$. Let h be the classifier which assigns $Z_i = h(t_i)$ to the test instance t_i as the predicted label set. Apart from producing multi-label prediction, a multi-label learning system outputs a real valued function of the form $f : T \times \gamma \to \Re$. For an test instance ($t_i$, Y_i), an ideal learning system output larger values for labels in Y_i than those not in Y_i, i.e., $f(t_i, y) > f(t_i, y')$ for any $y \in Y_i$ and $y \notin Y_i$. A ranking function $\psi(.,.)$ maps the output of $f(t_i, y)$ for any $y \in \gamma$ to $\{1, 2, \ldots, |\gamma|\}$ such that if $f(t_i, y_1) > f(t_i, y_2)$ then $\psi(t_i, y_1) < \psi(t_i, y_2)$.

The *Hamming Loss* (HL) evaluates how many disagreements are there in the actual and predicted label sets for a test instance and is defined as

$$HL = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{|Y_i \Delta Z_i|}{|\gamma|}$$

One Error (OE) gives the measure of how many times the top-ranked label is not in the actual label set for an instance. The OE value of 0 indicates a perfect performance; the smaller the value of OE, the better the performance.

$$OE(f) = \frac{1}{|T|}\sum_{i=1}^{|T|}\Lambda(\arg\max_{y\in\gamma} f(t_i, y) \notin Y_i)$$

where $\Lambda(p)$ equals 1 if the predicate p is true and 0 otherwise.

The *Coverage* (COV) indicates how far it is needed, on the average, to go down the list of labels to cover all the actual labels for an instance. The smaller the value of COV, the better the performance.

$$COV(f) = \frac{1}{|T|}\sum_{i=1}^{|T|}\max_{y\in Y_i}\psi(t_i, y) - 1$$

The *Average Precision* (AvP) evaluates the average fraction of labels ranked above a particular label $y \in Y_i$ that actually are in Y_i. The value of 1 for *AvP* indicates the perfect performance; the bigger the value, the better the performance.

$$AvP(f) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{1}{|Y_i|}\sum_{y\in Y_i}\frac{|\{y' | \psi(t_i, y') \le \psi(t_i, y), y' \in Y_i\}|}{\psi(t_i, y)}$$

IV. EMOTION CORPUS AND AGGREGATION

Emotional framing [22, 23] in news media is a very popular idea for writing emotionally charged

news articles. Each news item is shaped into a form of story with layered dramatic frames, e.g., fear caused by danger; sorrow and grief arising from violence, crime and death; exhilaration and tearful joy resulting from good luck or victory. Due to its effectiveness, the idea of emotional framing has been adopted universally by all news publishing houses. As a consequence, amount of news articles capable of evoking emotions in readers is huge. In this study, we perform reader emotion analysis on sentences extracted from news articles.

The emotion text corpus collected by us consists of 1000 sentences extracted from Times of India news paper archive (http://timesofindia.indiatimes.com/archive.cms). The sentences were collected from headlines as well as bodies of articles belonging to political, social, sports and entertainment domain.

Data Set	Anger	Disgust	Fear	Happiness	Sadness	Surprise	LD	LC
ORG	60	301	238	311	263	69	1.26	0.18
AD-COMB	-	350	238	311	263	69	1.25	0.21
ADC-COMB-SR	-	350	283	311	263	-	1.23	0.22

Corpus statistics. (ORG = Original AD-COMB = Anger Disgust combined, SR = Surprise removed)

4.1. Corpus Annotation

As discussed earlier, the reader of the sentence is experiencing a set of emotions. For example, in the following sentence, the reader may experience *fear* and *sadness* emotion.

'A double bomb attack in the town of Lakhdaria killed 12 people.'

The corpus annotation was performed by multiple annotators. The annotation scheme considers the following points:

- *Choice of emotion classes*: The annotation scheme considers six basic emotions, namely, Anger, Disgust, Fear, Happiness, Sadness, Surprise as specified by Ekman [24].
- *Multi-label annotation*: A sentence may trigger multiple emotions simultaneously. So, one annotator may classify a sentence to more than one emotion categories.

4.2. Agreement Study

As emotion is a subjective entity, readers' opinions about a sentence may vary. Thus it is of prime requirement to measure agreement in emotion corpus annotation and generating aggregated data by consulting the annotations by multiple annotators. The data was annotated by 5 annotators. The multi-label agreement coefficient Am [25] has been used to measure the agreement of emotion corpus annotation. The agreement value was computed to be 0.754 ($p < 0.0001$). In further analysis following observations were made.

- The (disgust, anger) pair is the most confusing pair. One psychological explanation of this observation is that disgust is of two types: core disgust and socio-moral disgust. Socio-moral context is dominant in news domain as it reports several facts of sexual harassment, law breaking, crime etc. It has been observed that anger is very co-associated with socio-moral disgust [26]. Combining the anger and disgust class to socio-moral disgust class improves the agreement value to 0.798.
- Surprise is one of the most ambiguous pair after combining anger and disgust together. This may be attributed to the fact that it may belong to positive as well as negative emotion category. Furthermore, evaluation of this emotion depends largely on background knowledge of the readers as compared to the other emotions. Removing this emotion from the corpus improves the agreement value to 0.815.

In section 5.2, we shall be investigating whether the machine classification of emotion is correlated with the human annotation behavior reported above. Majority voting is a popular technique for aggregation where, to attach a set of labels to a data item in the aggregated data, the majority decision is considered. We adopted this technique for generating aggregated corpus. The distribution of the sentences among the emotion categories, label density (LD) and label cardinality (LC) [10] in the original corpus as well as the modified corpus is given in Table 1.

V. EMOTION CLASSIFICATION

Emotion classification is performed with ADTboost.MH algorithm implemented in adtree package. In this section, we describe the features used for emotion classification and performance of the system.

5.1. Feature Extraction

Two types of features have been considered in our work: word feature and polarity feature.

Word Feature

Words sometimes are indicative of the emotion class of a text segment. For example, the word `bomb' may be highly co-associated with fear emotion. Thus words present in the sentences are considered as features. Before creating the word feature vectors, following preprocessing steps are adopted.

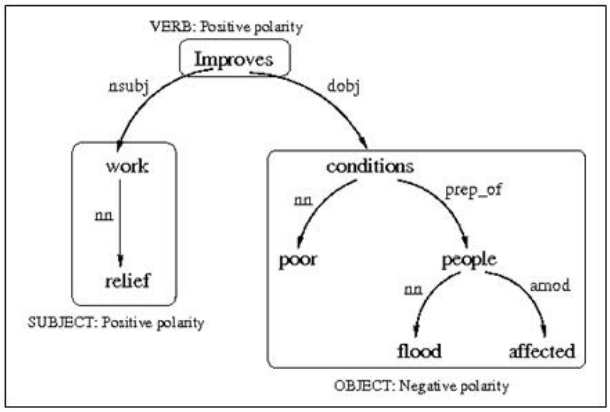
Stop words are removed.

Named Entities may introduce potential noise in emotion classification. So, named entities are removed using the Stanford Named Entity Recognizer [27].

The remaining content words are stemmed using Porter's stemmer algorithm [28].

1. nn(work-2, Relief-1)
2. nsubj(improves-3, work-2)
3. amod(conditions-5, poor-4)
4. dobj(improves-3, conditions-5)
5. nn(people-9, flood-7)
6. amod(people-9, affected-8)
7. prep_of(conditions-5, people-9)

(a). Dependency relations



(b). Dependency parse tree

Figure 4. Dependency parse output for example sentence

Polarity Feature

Polarity of the subject, object and verb of a sentence may be good indicators of the emotions evoked. For example, let us consider the following sentence.

Relief work improves the poor conditions of flood affected people.

Here, the subject, Relief work, is of positive polarity; the verb, improves, is of positive polarity; and the object phrase, poor conditions of flood affected people, is of negative polarity. Intuitively, a positive subject performs a positive action on a negative object and this pattern evokes a positive emotion.

Words in the sentence are tagged manually with their polarities. The example sentence presented above is tagged in the following manner (P = positive, N = negative, unmarked words are considered to be neutral).

[Relief]/P work [improves]/P the [poor]/N conditions of [flood]/N affected people.

The extraction of subject, verb and object phrase polarities in a sentence involves the following steps.

- *Extraction of subject, object and verb phrases (SOV Extraction):* The Stanford Parser [29] is used to parse the sentences and the subject, verb and object phrases are extracted by considering the dependency relations (nsubj, dobj, etc.) obtained as parser output. For the example sentence, relations and parse tree obtained as parser output are presented in Fig 1.

The subject phrase is extracted from relations 1 and 2. First argument (improves) of relation 2 is the verb in the sentence and second argument (work) is the head word of the subject phrase. The modifier (Relief) to the head word of the subject phrase is obtained from relation 1. The head word (conditions) of the objective phrase is obtained from relation 4, whereas the modifier (poor) is extracted from relation 3. We ignore the prepositional phrases in the current study. The output of the SOV extractor is as follows.

SUBJECT: relief works
VERB: improves
OBJECT: poor conditions

- *Polarity Assignment:* The polarity assignments to the phrases are performed

1. IF ((head-word is positive) AND (modifier is positive)) THEN phrase-polarity is positive
Example: [great]/P [win]/P --> [great win]/P
2. IF (head-word is negative) THEN phrase-polarity is negative
Example: airplane [hijack]/Ne --> [airplane hijack]/Ne
3. IF ((head-word is neutral) AND (modifier is positive) THEN phrase-polarity is positive
Example: [excellent]/P performance --> [excellent performance]/P
4. IF ((head-word is neutral) AND (modifier is negative)) phrase-polarity is negative
Example: [bad]/Ne dream --> [bad dream]/Ne
5. IF ((head-word is neutral) AND (modifier is neutral)) THEN phrase-polarity is neutral
Example: Minor girl --> Minor girl
6. IF (modifier is null) THEN phrase-polarity is the head-word polarity
Example: [kill]/Ne --> [kill]/Ne

with a set of rules. Some of the rules are given below.

Applying the rules on the example sentence the extracted subject, object and verb polarities are as given below.

SUBJECT: [relief works]/P (Rule 3)
VERB: [improves]/P (Rule 6)
OBJECT: [poor conditions]/Ne (Rule 4)

5.2. Experimental Results

In this section, we report results of applying ADTboost.MH in three different versions of emotion corpus: original corpus (ORG); anger-disgust combined corpus (AD-COMB); anger-disgust combined and surprise removed corpus (AD-COMB-SR). For each corpus, we perform 5 fold cross validation and for each trial the number of boosting rounds are selected by observing the variation in average precision over the number of boosting steps. The start of the saturation point is taken to be the number of boosting steps (see Fig 2). Fig 6. presents (Appendix A) the ADTree generated in training with word and polarity feature combination.

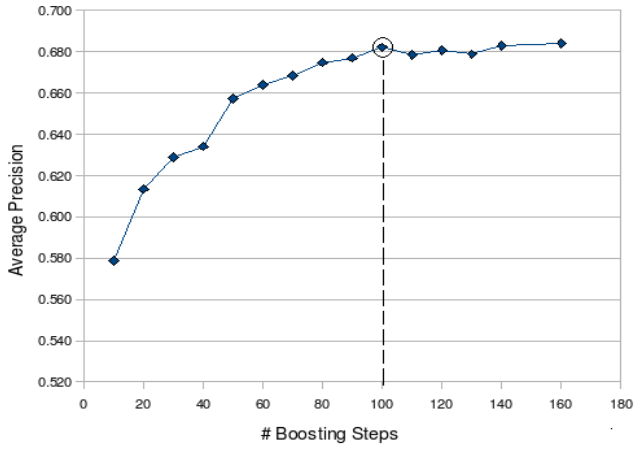


Figure 5. Average precision Vs. number of boosting steps for word feature model. The saturation in this curve starts at boosting step of 100. So, the number of boosting steps for the trial is set to be 100.

Comparison of Features

We have used two types of features for emotion classification: word feature and polarity of subject, verb and object feature. For three different experiments, the comparisons of features are provided in Fig 3.

It may be observed that OE, HL and COV values in all the experiments are smaller for feature set consisting of word and polarity (word+polarity set) as compared to the feature set that consists of only word (word set). Further, the value of average precision is higher for word+polarity set as compared to word set. Thus it may be conjectured that the performance of emotion classifier is improved with the inclusion of polarity of subject, verb and object information.

Human Classification Vs. Machine Classification

In this section, we shall be investigating whether the machine classification of emotion correlates with the emotion classification by human. In agreement measurement of emotion corpus (section 4), three different corpora ORG, AD-COMB and AD-COMB-SR can be ordered according to the agreement values in the following way.

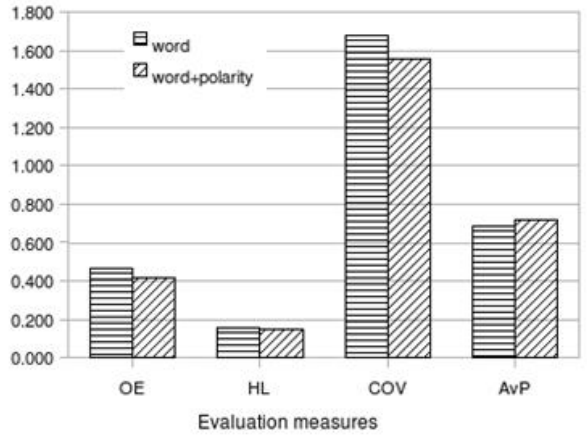
AD-COMB-SR » AD-COMB » ORG

where A » B denotes that human annotators agree better on corpus A than B. As the word and polarity feature combination provides the best results in different versions of the corpus, we report results on different multi-label evaluation measures for this combination only. Table 2 presents the results on each evaluation criteria with the best result shown in bold face.

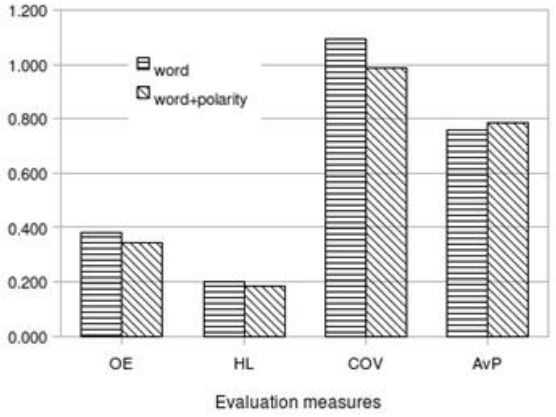
Evaluation Criteria	ORG	AD-COMB	AD-COMB-SR
One Error	0.414	0.342	**0.336**
Hamming Loss	**0.146**	0.200	0.216
Coverage	1.554	0.986	**0.838**
Average Precision	0.714	0.786	**0.795**
Micro-F1	34.93	46.30	**59.52**

TABLE 2. EXPERIMENTAL RESULTS ON MULTI-LABEL EMOTION CLASSIFICATION FOR THREE DIFFERENT DATA SETS

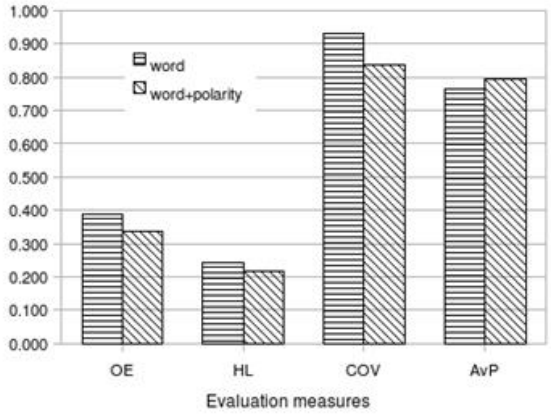
To compare the results in three different data sets, a partial order ">" is defined on the set of all



(a) Experiment 1: original corpus



(b) Experiment 2: anger-disgust combined corpus



(c) Experiment 3: anger-disgust combined with surprise removed corpus

Figure 3. Comparison of features in three experiments (OE = one error, HL = hamming loss, COV = coverage and AvP = Average precision)

comparing corpus for each evaluation criterion, where C1 > C2 means that the performance of the emotion classifier is better in C1 than C2 on the specific metric. Relative performance of emotion classification in three different data sets is provided in Table 3. To measure the overall performance in a particular corpus, a score is assigned to it which takes into account the relative performance in other corpus on all metrics, i.e., if C1 > C2, C1 is rewarded with a positive score +1 and C2 is penalized with -1. Based on the accumulated score, a total order "»" is defined where C1 » C2 denotes that the performance of emotion classifier is better in corpus C1 than corpus C2. The comparison result is shown in Table 3.

Evaluation Criteria	Ordering of Corpus
One Error	C1 >C2, C1>C3, C2 >C3
Hamming Loss	C2 >C1, C3 >C2, C3>C1
Coverage	C1 >C2, C1 >C3, C2>C3
Average Precision	C1 >C2, C1>C3, C2 >C3
Micro-F1	C1 >C2, C1>C3, C2 >C3
Total Order	C1(6) » C2(0) » C3(-6)

TABLE 3. RELATIVE PERFORMANCE OF EMOTION CLASSIFICATION IN DIFFERENT DATA SETS (C1 = AD-COMB-SR; C2 = AD-COMB; C3 = ORG).

From Table 3, it may be noted that machine classification of emotion exhibits the same relative performance on different data sets as observed in human classification of emotion.

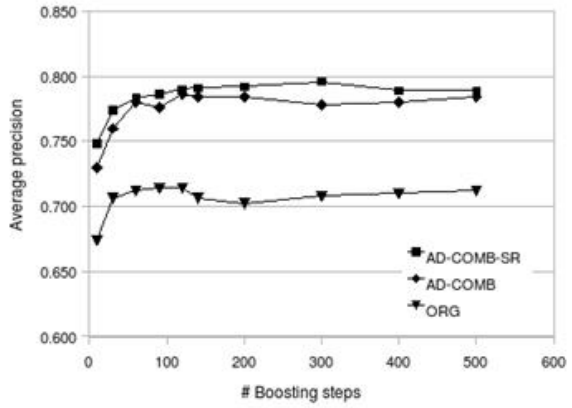
Other Results

Here we present the variation of different multi-label evaluation metrics with number of boosting steps in Fig 4. In all the metrics except hamming loss, classification of sentences into emotion classes shows the following relative performance pattern.

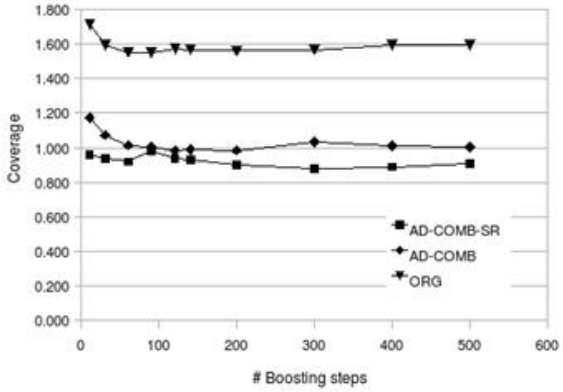
AD-COMB-SR > AD-COMB > ORG

The performance in AD-COMB-SR and AD-COMB corpus are far better than that in ORG corpus. One exception to the above relative performance trend is the hamming loss metric. The relative performance follows the pattern given below.

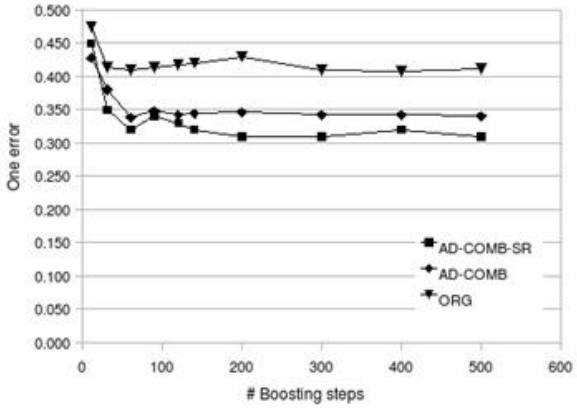
ORG > AD-COMB > AD-COMB-SR



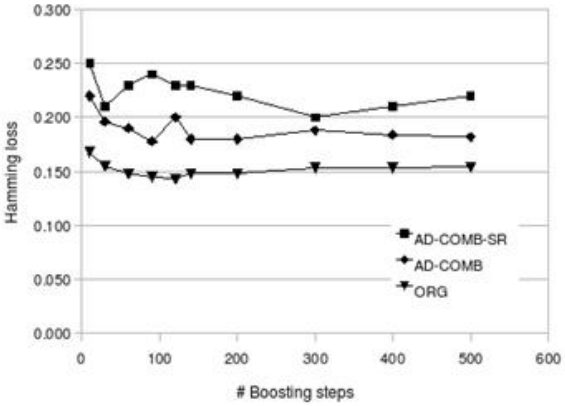
(a) Average precision Vs. number of boosting steps curve



(b) Coverage Vs. number of boosting steps curve



(c) One error Vs. number of boosting steps curve



(d) Hamming loss Vs. number of boosting steps curve

Figure 4. Variation of multi-label measures with number of boosting steps

In terms of average predicted label per data item (sentence) the following pattern is observed.

AD-COMB-SR (1.13) > AD-COMB (0.816) > ORG (0.542)

Higher the average predicted label, higher is the chance of mismatch in the predicted and actual vector of labels and the value of Hamming loss is higher if the such cases of mismatches are higher. This explains the exception in performance trend in terms of Hamming loss.

VI. DISCUSSIONS

In this paper, we have presented a multi-label emotion classification model based on ADTboost.MH algorithm. Two types of features, namely, word and polarity of the subject, object and verb have been used. The experiment on comparison of features revealed that word features combined with the polarity feature provides better performance than the feature set consisting only of words. In the next set of experiments, the same performance pattern has been observed in both human and machine classification of emotion. The best performance was observed in the AD-COMB-SR corpus with word and polarity as the feature (AvP=0.795, OE=0.336, HL = 0.216, COV=0.838 and Micro average F1=59.52%).

Emotional framing of the news articles is not demographically invariant. For example, Chinese and U.S. media gave different explanations to April 1st Military Airplane Collision [30]. These variations may result in differences in the emotional responses of the readers belonging to demographically distant locations (e.g., two different countries). In this work, we have not addressed these differences in emotional framing and considered the annotators to belong to the same demographic location and homogeneous social background (e.g., education, economy etc.)

Fig 5 presents the word feature count vs. word rank curve. The curve suggests that word feature count follows the Zipfian distribution (power law fit with $R^2 = 0:93$)..

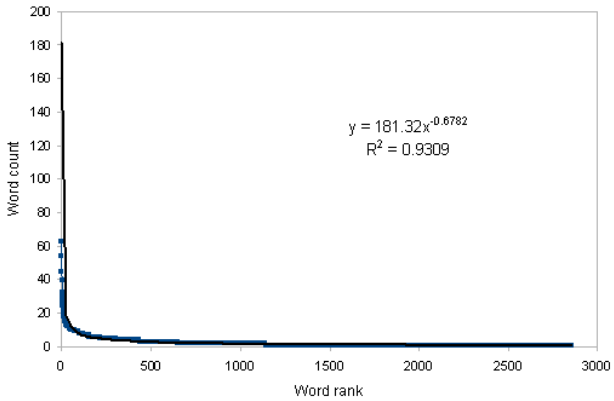


Figure 5. Distribution of word features

It has been observed that approximately 10% of the total word features have counts five or above. Thus feature sparseness problem is one issue that needs to be handled for improving classification performance. Generalization of features is one of the techniques to handle feature sparseness problem. In future study, we shall be investigating on employing different generalization techniques in emotion classification.

APPENDIX A

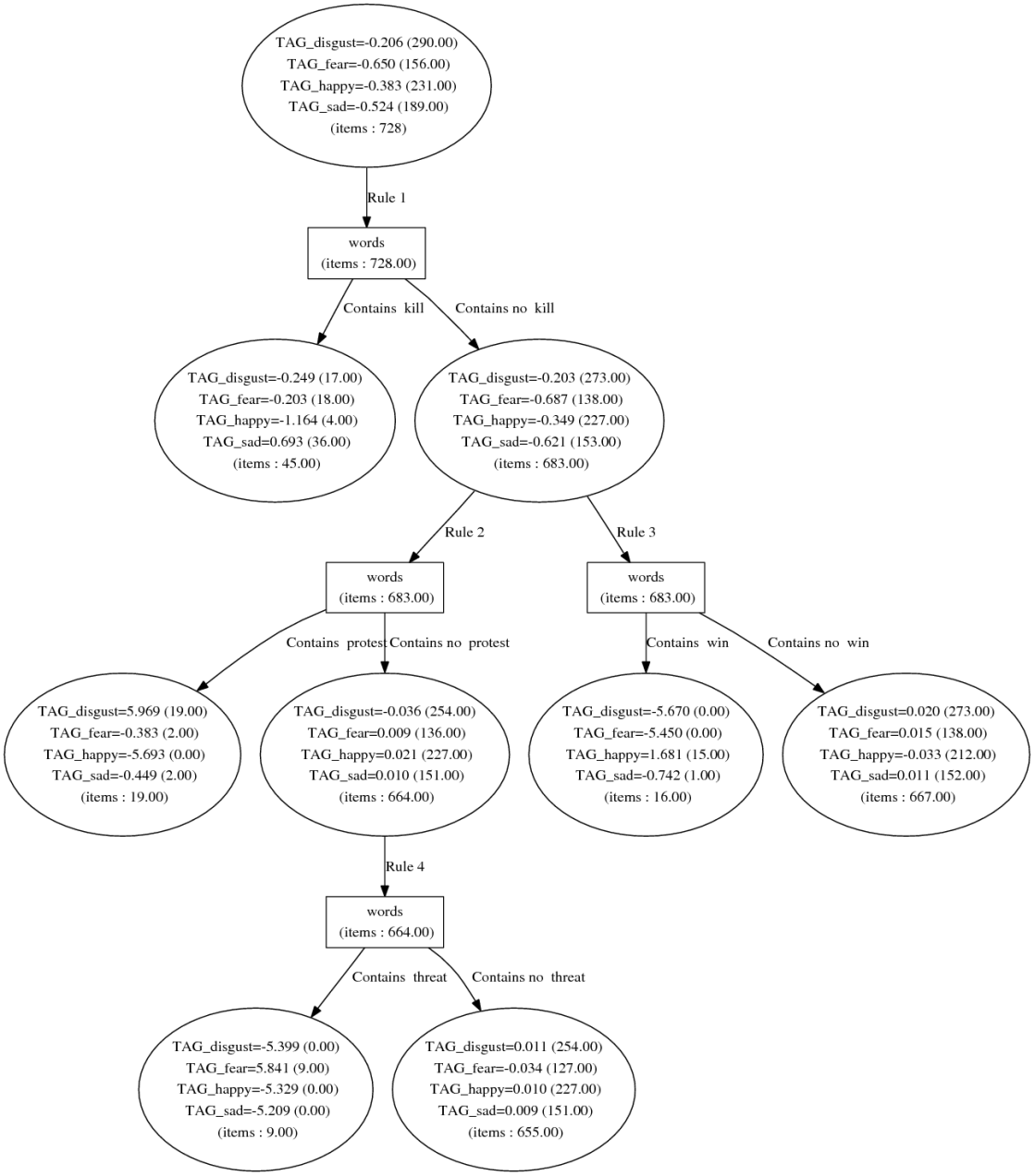


Figure 6. ADTree generated in training with word feature with number of boosting steps = **5.**

REFERENCES

[1] Pantic, M., Rothkrantz, L.J.M. (2000). Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12) p. 1424-1445.

[2] Nwe, T.L., Foo, S.W., De Silva, L.C. (2003) Speech emotion recognition using hidden markov models. Speech Communication 41(4) p. 603-623.

[3] Kovecses, Z. (2003) Language and emotion concepts. In: Metaphor and Emotion: Language, Culture, and Body in Human Feeling. Cambridge University Press, Cambridge.

[4] Subasic, P., Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. IEEE Transaction on Fuzzy Systems 9(4) p. 483-496.

[5] Grefenstette, G., Qu, Y., Evans, D., Shanahan, J. (2006). Validating the coverage of lexical resources for a_ect analysis and automatically classifying new words along semantic axes. p. 93-107.

[6] Mihalcea, R., Liu, H. (2006). A corpus-based approach to finding happiness. In: In AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs, AAAI Press. p. 139-144.

[7] Mishne, G. (2005). Experiments with mood classification in blog posts. In: Proceedings of the 1st Workshop on Stylistic Analysis of Text for Information Access.

[8] Wu, C.H., Chuang, Z.J., Lin, Y.C. (2006). Emotion recognition from text using semantic labels and separable mixture models. ACM Transactions on Asian Language Information Processing (TALIP) 5(2) p. 165-183.

[9] Abbasi, A., Chen, H., Thoms, S., Fu, T. (2008). Affect analysis of web forums and blogs using correlation ensembles. IEEE Transactions on Knowledge and Data Engineering 20(9) p. 1168-1180.

[10] Tsoumakas, G., Katakis, I. (2007). Multi label classification: An overview. International Journal of Data Warehouse and Mining 3(3) p. 1-13.

[11] De Comite, F., Gilleron, R., Tommasi, M. (2003). Learning multi-label alternating decision trees from texts and data. Machine Learning and Data Mining in Pattern Recognition. p. 251-274.

[12] Strapparava, C., Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In: Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval2007), Prague, Czech Republic.

[13] Chaumartin, F.R. (2007). Upar7: A knowledge-based system for headline sentiment tagging. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, Association for Computational Linguistics. p. 422-425.

[14] Fellbaum, C. (1998). WordNet An Electronic Lexical Database. The MIT Press, Cambridge, MA ; London

[15] Kozareva, Z., Navarro, B., Vazquez, S., Montoyo, A. (2007). Ua-zbsa: A headline emotion classi_cation through web information. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, Association for Computational Linguistics. p. 334-337.

[16] Katz, P., Singleton, M., Wicentowski, R. (2007). Swat-mp: The semeval-2007 systems for task 5 and task 14. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, Association for computational Linguistics. p. 308-313

[17] Roget's new millennium thesaurus, 1st ed. Lexico Publishing Group, LLC (2007)

[18] Lin, K.H.Y., Yang, C., Chen, H.H. (2008). Emotion classification of online news articles from the reader's perspective. In: Web Intelligence. p. 220-226.

[19] Lin, K.H.Y., Chen, H.H. (2008). Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, Association for Computational Linguistics. p. 136-144.

[20] Freund, Y., Mason, L. (1999). The alternating decision tree learning algorithm. In: ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. p. 124-133.

[21] Schapire, R.E., Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. Machine Learning 39(2/3) p. 135-168.

[22] Corcoran, P.E. (2006). Emotional framing in australian journalism. In: Australian & New Zealand Communication Association International Conference, Adelaide, Australia, ANZCA

[23] Go_man, E. (1986) Frame Analysis: An Essay on the Organization of Experience. North-eastern University Press

[24] Ekman, P., Friesen, W.V., Ellsworth, P. (1982). What emotion categories or dimensions can observers judge from facial behavior? In Ekman, P., ed.: Emotion in the human face. Cambridge University Press, New York p. 39-55.

[25] Bhowmick, P.K., Basu, A., Mitra, P. (2008). An agreement measure for determining inter-annotator reliability of human judgements on affective text. In: Coling 2008:Proceedings of the workshop on Human Judgements in Computational Linguistics, Manchester, UK, Coling 2008 Organizing Committee, 58-65

[26] Simpson, J., Carter, S., Anthony, S.H., Overton, P.G. (2006). Is disgust a homogeneous emotion? Motivation and Emotion 30(1) p. 31-41.

[27] Finkel, J.R., Grenager, T., Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics p. 363-370.

[28] Porter, M.F.: An algorithm for suffix stripping. Program (3) p. 130-137.

[29] Klein, D., Manning, C.D. (2003). Accurate unlexicalized parsing. In: ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics. p. 423-430.

[30] Peng, W. (2003). A textual analysis of the news framing of the april 1st military airplane collision by people's daily and the new york times. In: 2nd Hawaii International Conference on Social Sciences, Honolulu, Hawaii

[31] Esuli, A., Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In: In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06. p. 417-422.