Proceedings of the 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), 26-28 September, Dhaka, Bangladesh

A Computational Approach of Recognizing Emotion from Bengali Texts

Hasan Abid Ruposh and Mohammed Moshiul Hoque
Dept. of Computer Science & Engineering, Chittagong University of Engineering & Technology, Chittagong-4349, Bangladesh
{hasan.ruposh, mmoshiulh}@gmail.com

***Abstract*— Emotion recognition is the task of determining distinct emotion exhibited in text. In recent year, due to the availability of enormous amount of textual data, specially dogmatic and self expressiveness of text played a significant role to lead focus in this area. This paper presents an emotion recognition technique that can identify six basic emotions from Bengali texts such as happy, sad, anger, fear, surprise, and disgust respectively. We develop a corpus consisting of 1200 emotive words that are used to train the SVM classifier for identifying different emotions. Experimental result shows that the proposed system can recognize emotions with 73% accuracy which is higher than the Naive based approach (60%).**

Keywords–Bangla language processing; Emotion recognition; Feature extraction; Emotion corpus; Evaluation

I. INTRODUCTION

Emotion involves experience, cognition, feelings, behaviour, physiology, and conceptualization [1]. There are numerous approaches are employed to identify emotions from the humans such as body gesture, facial expressions, heart rate, blood pressure, and text information. In this work, we pay attention on the recognition of emotion from Bengali texts. Emotion recognition from text is a growing research topic of NLP/computational linguistics that is intently related to sentiment analysis. Sentiment analysis focuses to identify negative, positive, or neutral feelings from text, whereas emotion analysis focuses to identify kinds of feelings through the text expression, such as happy, disgust, anger, sad, surprise and fear.

Recently, emotion recognition in text has become more attractive topic due to its broad utilization in marketing, psychology, HCI, advertising, artificial intelligence, pervasive computing, etc. Therefore, interpretation of emotions may advantageous to any company or in such as private enterprise, managing the response to a typical disaster, calculating happiness index, developing better interactive AI agents, analyzing consumer reaction, assessing the impact of the products on particular population, recommendation system, question-answering system and so on [2]. There are six basic emotions that a human is capable of expressing through facial expression: happiness, sadness, anger, fear, surprise, and disgust [3]. Human express their emotions through their facial expressions, speech, body gestures or writings. Various distinct sources of information, such as text, speech, and visual can be considered to analyze humans' emotions. In this work, we focus on the recognition of six basic emotions from Bengali text only.

Emotion recognition in text refers to the use of computational linguistic, and natural language processing to determine discrete emotional information from the source texts. Text is the most common form of interaction medium on the web and Recently, due to the rapid growth of Internet usages and evolution of web 2.0, users are updating huge amount of text contents on the Web in the form of social media posts, micro-blogs, news articles, etc. These contents can be used to develop better interactive system which needs to be able to analyze the text and deduce the emotion of the end user.

Determining emotions from the texts is a quite challenging and complicated task due to the evasive characteristic of emotion expression in text and also the intricacy of human emotions [2]. Although Bengali is the sixth-most widely spoken language in the world there is no usable computational system is developed that can recognize emotions from Bengali texts. A significant research activities have been carried out on emotion recognition in text, especially, in English and European languages. However, a very few work are done on sentiment analysis in Bengali text [4][5]. In addition, there is no useful work has been conducted yet to recognize the six basic emotions from Bengali texts in Bangladesh. Thus, in this work we proposed a computational technique of recognizing emotions from Bengali text using machine learning algorithm.

II. RELATED WORK

There is a significant number of work have been conducted in English, Chinese or European languages to detect emotion from the text data which are broadly categories into three approaches: keyword based [6][6], learning based [7] and hybrid based [8]. These work used the features that were adopted from semantic and syntactic data to detect emotions. Several work have been done using hashtags as the emotional label for the data and SVM as the classifier. Purver et al. [9] used SVM classifier on Twitter data and gained 82% accuracy for categorizing the emotion Happy, and 67% in categorizing over the whole dataset for the identical emotion. Balabantaray et al. [10] conduced an emotion classification task on Twitter data in which 8000 tweets are labelled manually for six basic emotions. This work used multi-class SVM with 73.24% accuracy. A study was carried out by Seyeditabari et al. [2] for classifying comments in social media. An unsupervised method was proposed to automatically identify emotions in text, based on

978-1-7281-4934-9/19/$31.00 ©2019 IEEE

categorical as well as dimensional approaches of emotions [11]. An automatic tweets based emotion detection system is developed by Hasan et al. [12].

Emotion detection from Bengali text is a relatively quite new research issue in Bangla language processing field. Few attempts are made to classify sentiment from Bengali text into positive, negative or neutral categories. Shaika et al. [13] presented a methodology to extract the sentiment into positive or negative category from Twitter posts. In order to classify the posts they have used SVM and Maximum Entropy algorithms. A Naive bayes approach is developed to classify sentiment into positive, negative or neutral from both English and Bangla texts [14]. This method used the Amazon reviews as data sets and achieved the 85.7% and 85.0% accuracy for English and Bengali review texts respectively. Islam et al. [15] described a supervised approach based on Nave bayes to recognizing sentiment into positive or negative classes. They have used Facebook status written in Bengali as source data sets and achieved 72.0% accuracy. A deep learning based sentiment detection is proposed by Hayder et al. [5]. This method classify the sentiment into three categories: negative, positive, neutral emotion and gained 78.0% accuracy. A method is proposed based on TF.IDF algorithm to detect sentiment from Bengali social media texts which detected positive, negative and neutral sentiment categories [16]. Das et al. [17] is conducted a study to identify emotions in Bengali blog texts. They used rule based baseline system and SVM to detect emotional expression from blogs. An another work of emotion analysis based on conditional random field is proposed that detected six basic emotions: sad, happy, fear, anger, surprise and disgust [18]. In this work classification is done on Bengali blogs and News texts at word and sentence level.

Most of the previous studies focused on the sentiment detection in terms of positive, negative or neutral categories from Bengali blogs, tweets or social media texts. In the proposed approach, our main task is to recognize the basic six emotions such as happy, disgust, fear, surprise, fear and anger respectively from Bengali texts.

III. Proposed Methodology

Fig.1 illustrates the schematic representation of proposed approach of emotion recognition. This approach composes of four major modules: training, classification, testing and recognition.

A. Data Preprocessing

In order to fill in missing values, smooth noisy data, and resolve inconsistencies data prepossessing is needed.

1) Tokenization: Tokenization divides the sentences into individual words. Sentences may be broken into distinct words and punctuation across the white spaces. Fig.2 shows the results of tokenization of a text input 'ami besh bhalo achi'.

2) Append back remaining words in the text: After tokenization process, rest of the texts are appended back in the

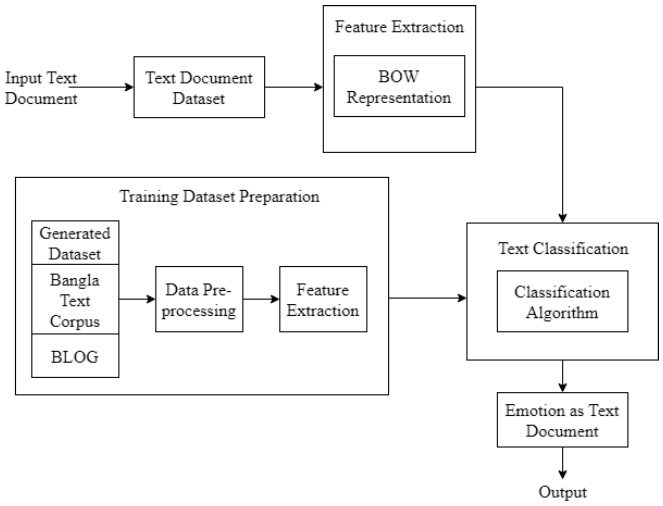

Fig. 1. Proposed approach of emotion recognition in Bengali text

Index	Type	Size	Value
0	str	1	আমি
1	str	1	বেশ
2	str	1	ভালো
3	str	1	আছি

Fig. 2. Tokenized text

text and these texts are included in the corpus. Fig.3 shows a fragment of the corpus after cleaning.

Insignificance or irrelevant words are removed from the text. We use main bodies of the text to train the emotional text classifier and represent the text document using a list of words and their frequencies. Finally, manually tagged those data into six emotion categories.

B. Feature Extraction

Word frequencies are used as features quite often to extract the feature from the text.

1) CountVectorizer: CountVectorizers used to learn the vocabulary of a set of texts and then transform them into a data-frame that can be used for building models. CountVectorizer take few parameters that are important for extracting features.

- **Max-Features:** Most frequent 500 words are used to reduce time and storage complexity. This is used to minimize the sparse matrix.
- **Stop Words:** In the proposed system, we used those words into account as the dataset is smaller in size which improves system accuracy.

2) Bag of Words: A bag-of-words express the distribution of words within a text document. It take account of two factors: a vocabulary of known words and a count of the existence of known words. In bag-of-word model, histogram of the words within text is investigated in which each word count is considered as a feature. Each text document is converted into binary vector that may use as input or output

Index	Type	Size	Value
489	str	1	আমার দেশপ্রেমীদের শান্তিপূর্ণ জীবনকে হুমকির মুখে ফেলার অপরাধে ক্রমবর্ধ ...
490	str	1	যেভাবে বিধবাদের আচরণ করা হয়েছে সেটি ন্যক্কারজনক বিধবাদেরকে বারবার উত্ ...
491	str	1	মানুষের জন্য প্রাথমিক প্রয়োজন হচ্ছে পানি এই পানি বাণিজ্য হিসাবে বিক্র ...
492	str	1	আজকে এক জরুরী মিটিংয়ের জন্য গাড়িতে চড়তেই দেখি গাড়ি স্টার্ট নিচ্ছেনা ধুর ...
493	str	1	তারা বিপক্ষের কাছ থেকে অবজ্ঞাজনক প্রস্তাব পেয়ে সেখান থেকে প্রস্থান করল ...
494	str	1	নতুন মেট্রো রেল পরিবহণ একটি লজ্জাজনক জঘণ্য অপ্রীতিকর যোগাযোগ মাধ্যম এত ...
495	str	1	এই ভবনের অবস্থা বর্তমান দিনের স্যানিটেশনের আলোকে ঘৃণ্য এবং নিঃসন্দেহে ...
496	str	1	তারা ভেবেছিল তারা কাজটি করতে পারবে কিন্তু অপ্রীতিকর রেজাল্ট দেখে তারা ...
497	str	1	এদের বিরুদ্ধে একটা প্রতিবাদ দরকার এদের গাছের সাথে বেঁধে পেটানো উচিত ডি ...
498	str	1	আপনাদের লজ্জা করে না এক নম্বর নেটওয়ার্ক দাবি করতে ঢাকা শহরে ঠিক মত নেটওয় ...

Fig. 3. A fragment of developed Bengali corpus

in a learning model. Histogram intensity of the word is calculated as in Eq.1 [19].

$$I = \frac{N_{key}}{N_{total}} \tag{1}$$

Where, I denotes the intensity, N_{key} denotes number of keywords in a emotion text and N_{total} denotes the total number of words in the text respectively.

Feature space is a two dimensional array where rows represents each text of the corpus and columns represents number of unique words available in the corpus. Each cell of the array represents the number of time a specific word occurs in a specific text. Fig. 4 shows a small fragment of training set feature space.

	1	2	3	4
297	0	0	0	0
298	0	0	0	0
299	0	1	0	0
300	0	0	0	0
301	0	0	0	0
302	0	1	0	0
303	0	2	0	0
304	0	0	0	0
305	0	0	0	0

Fig. 4. A fragment of feature space of training set

3) Predicted Level: In the proposed implementation, predicted level (i.e., emotional categories) of test sample is labeled 0, 1, 2, 3, 4, or 5 [depending on the categories]. It represents the semantic interpretation of text in the test sample. If the sample text is not labeled within the categories then the system will fails to process it. Fig. 5 shows a sample output of the predicted level of text sample.

C. Classifier

All extracting features are used to train the classifier model. We used SVM classifier with linear and non-linear classification using kernel trick. For linear classification, cost

Index	Type	Size	Value
0	str	1	1. Happiness
1	str	1	2. Sadness
2	str	1	3. Anger
3	str	1	4. Fear
4	str	1	5. Surprise
5	str	1	6. Disgust

Fig. 5. A sample predicted level of texts

function can be evaluated by the Eq. 2.

$$\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(w.x_i - b)\right) + \lambda(\|\mathbf{w}\|)^2 \tag{2}$$

where, λ identifies the trade-off between the margin-size and confirming that the samples remain on the actual side of the margin. Therefore, the 2nd term in the loss function become insignificant for smaller values of λ which behaves like as hard-margin SVM. General linear kernel trick is determined according to Eq. 3.

$$k(x_i, x_j) = (x_i \cdot x_j)^d \tag{3}$$

D. Training and Testing Phases

A set of text file are used as training sample to train the classifier. Output of the training phase is a trained machine learning model that will be used in testing phase for emotion detection and recognition. A sample text is processed with tokenizer, and extracted necessary features which are used to learn the classifier model. Input to the testing phase is a text for which emotional categories is unknown or to be determined. Classifier module used extracted features to determine the emotion category for the test text sample.

IV. EXPERIMENTS

It is a quite challenging task to develop an useful system in Bangla language processing field due to the scarcities of available resources in Bengali. Due to the unavailability of emotion corpus in Bengali, we first focus to develop an emotion corpus to serve our purpose. We evaluate the proposed system in terms several standard matrices such as confusion matrix, precision, recall, F1 score and ROC measures respectively.

A. Corpus Preparation

Corpus contains 1200 emotional subjective Bengali text samples which are labeled in terms of six basic emotions. We collect half of our data from Cambridge English Corpus by translating them from English to Bengali using Google translator. Some of the emotional texts are collected from online blogs, Facebook pages, Bengali newspapers. In order to classify the emotion we used Ekmans basic emotion categories such as sad, happy, fear, anger, disgust and surprise [20]. We adopted the following properties of the text to labelling it into one of the emotion categories [21].

- A text is considered happiness if it contain emotional words of feeling well, showing joy or pleasure.
- A text is considered sadness if it contain emotional words of being affected with or expressive of grief or upset or failure.
- A text is considered anger if it contain emotional words that highly contrast or disagree with the emotion happiness or showing rage to someone.
- A text is considered fear if it contain emotional words that expresses an disagreeable emotion caused by the threat of endangerment, pain or harm.
- A text is considered disgust if it contain emotional words to offend the good taste or moral sense, extreme detest.

TableI summarizes the statistics of developed corpus.

TABLE I
SUMMARY OF THE BENGALI EMOTION CORPUS

Number of documents	1200
Number of sentences	3600
Number of words	12000
Total unique words	2137

1) Evaluation Measures: In order to evaluate our proposed system, we used several evaluation matrices such as confusion matrix, precision, recall, F1 score and ROC measures.

- **Confusion Matrix:** It is a tabular representation of data used for evaluating the classification model performance. As our system is a multi-class classification model, we used a confusion matrix consists of 7 (row) $\times$ 7 (column). This matrix represents total true positives, false positives, true negatives and false negatives numbers respectively.
- **Precision:** refers as positive predictive value. It calculates the ratio of exactly classified text into a particular class to the total number of classified texts of that emotion class. Precision can be obtained by Eq. 4

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

- **Recall:** calculates the ratio of correctly classified text into a particular class to the total number of classified texts of that emotion class (Eq. 5).

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

- **F1 score:** It is the weighted mean of recall and precision measures. Eq. 6 is used to calculate F1 score.

$$F_1 = \frac{2 \times recall \times precision}{recall + precision} \tag{6}$$

- **Accuracy:** Accuracy is used as a statistical evaluation of how well a classification test correctly determines or keep out a condition. Therefore, the accuracy is the proportion of true results both true positives and true negatives among the total number of test sample. Accuracy can be measured using the Eq. 7.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

B. Results

In order to measure the effectiveness, We used two classification algorithms: SVM and Naive Bayes. Table II shows the classification report for SVM with linear kernel. Fig. 6

TABLE II
PRECISION, RECALL AND F1 SCORE FOR EMOTION CLASSES USING SVM

	Precision	Recall	F_1 score	Support
Happiness	0.63	0.71	0.67	17
Sadness	0.50	0.43	0.46	14
Anger	0.65	0.69	0.67	16
Fear	0.90	1.00	0.95	19
Surprise	0.78	0.72	0.75	25
Disgust	0.81	0.76	0.79	17
avg./total	0.73	0.73	0.73	108

shows the ROC measures for the different emotion class of text using SVM. Table III shows the classification report

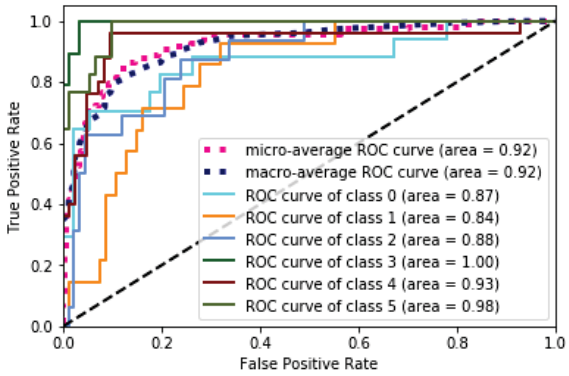

Fig. 6. ROC Curve (SVM Linear Kernel)

for Naive Bayes classifier. Here support represents the total number of document test by the system.

Table IV represents the comparison between SVM and Naive Bayes classification algorithms for all emotional classes in terms of accuracy. The result reveals that SVM is performed better than the Naive Bayes classifiers in recognizing emotions from Bengali texts. On average, SVM given 73% accuracy while Naive Bayes provided only 60%.

1) Sample Input-Output: Sample test texts (.txt file) are kept in a folder. These samples are processed by the proposed system which determines the corresponding emotion category of the test input. Fig. 7 depicts sample input texts and corresponding predicted level as emotion categories.

TABLE III

PRECISION, RECALL AND F1 SCORE FOR EMOTION CLASSES USING NAIVE BAYES

	Precision	Recall	F_1 score	Support
Happiness	0.60	0.53	0.56	17
Sadness	0.50	0.50	0.50	14
Anger	0.43	0.38	0.40	16
Fear	0.59	0.68	0.63	19
Surprise	0.78	0.72	0.75	25
Disgust	0.60	0.71	0.65	17
avg./total	0.73	0.73	0.73	108

TABLE IV

COMPARISON OF EACH EMOTION CLASS

Emotion Category	Naive Bayes (%)	SVM (%)
Happiness	0.53	0.75
Sadness	0.50	0.63
Anger	0.40	0.67
Fear	0.69	0.80
Surprise	0.72	0.75
Disgust	0.71	0.76

V. CONCLUSION

The main purpose of the proposed system is to classify the Bengali texts in terms of six basic emotions such as sadness, happiness, fear, anger, disgust, and surprise respectively. For this purpose, we developed a emotion corpus of Bengali text and trained SVM and Naive Bayes classifiers. The evaluation results shown that SVM is performed better than the Naive Bayes in terms of higher accuracy and lower error rate. The performance of the system can be improved with larger corpus including more emotive words.

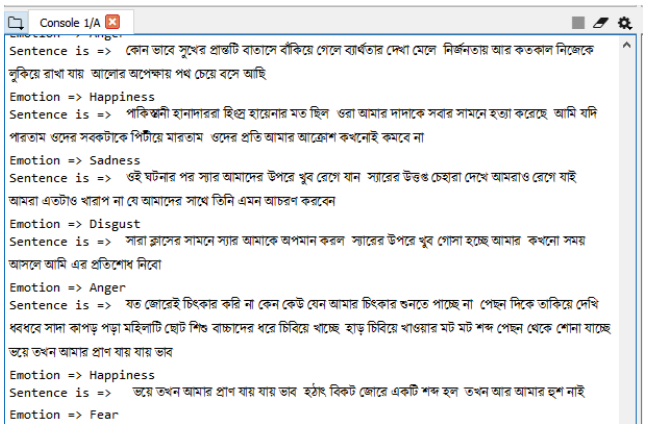

Fig. 7. Sample input and corresponding output

REFERENCES

[1] A. Ortony, G. Clore, and A. Collins, "The cognitive structure of emotions." 1988.

[2] A. Seyeditabari, N. Tabari, and W. Zadrozn, "Emotion detection in text: a review." *CoRR*, vol. abs/1806.00674, 2018.

[3] P. Ekman, "Facial expression and emotion." *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[4] D. Das and S. Bandyopadhyay, "Word to sentence level emotion tagging for bengali blogs." *ACL-IJCNLP*, pp. 149–152, 2009.

[5] M. S. Haydar, M. Al Helal, and S. A. Hossain, "Sentiment extraction from bangla text: A character level supervised recurrent neural network approach," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*. IEEE, 2018, pp. 1–4.

[6] T. Hancock, C. Landrigan, and C. Silver, "Expressing emotion in text-based communication." *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pp. 929–932, 2007.

[7] C. Yang, Y. Lin, and H. Chen, "Emotion classification using web blog corpora." *Proc. of IEEE/WIC/ACM Int. Conf. on Web Intelligence*, pp. 275–278, 2007.

[8] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text." *Proc. of Int. Conf. on Text, Speech and Dialogue*, vol. LNCS, 4629.

[9] M. Purver and S. Battersby, "Experimenting with distant supervision for emotion classication." *In Proc. of the 13th Conf. of the European Chapter of the Association for Computational Linguistics*, pp. 482–491, 2012.

[10] R. Balabantaray, M. Mohammad, and N. Sharma, "Multi-class twitter emotion classication: A new approach." *Int. J. of Applied Info. Sys.*, vol. 4, no. 1, pp. 48–53, 2012.

[11] S. Kim, A. Valitutti, and R. Calvo, "Evaluation of unsupervised emotion models to textual affect recognition." *In Proc. of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 62–70, 2010.

[12] M. Hasan, E. Rundensteiner, and E. Agul, "Automatic emotion detection in text streams by analyzing twitter data." *Int. J of Data Sci and Anal*, vol. 7, no. 1, pp. 35–51, 2019.

[13] C. Shaika and W. Chowdhury, "Sentiment analysis for bangla microblog posts," in *Proc. Int. Conf. on Informatics, Electronics and Vision*. IEEE, 2014.

[14] K. A. Hasan, M. S. Sabuj, and Z. Afrin, "Opinion mining using naive bayes," in *IEEE Int. WIE Conf. on Electrical and Computer Engineering*. IEEE, 2015, pp. 511–514.

[15] M. S. Islam, M. A. Islam, M. A. Hossain, and J. J. Dey, "Supervised approach of sentimentality extraction from bengali facebook status," in *Computer and Information Technology (ICCIT), 2016 19th International Conference on*. IEEE, 2016, pp. 383–387.

[16] M. Nabi, T. Altaf, and S. Ismail, "Detecting sentiment from bangla text using machine learning technique and feature analysis," *Int J of Com App*, vol. 153, no. 11, pp. 28–34, 2016.

[17] D. Das and S. Bandyopadhyay, "Emotions on bengali blog texts: role of holder and topic," in *Proc. of Int. Conf. on Advances in Social Networks Analysis and Mining*. IEEE, 2011, pp. 587–592.

[18] D. Dipankar and S. Bandyopadhyay, "Analyzing emotion in blog and news at word and sentence level," in *Proc. of the 4th Indian Int. Conf. on Artificial Intelligence*, 2009, pp. 1402–1414.

[19] S. Sriram, , and X. Yuan, "An enhanced approach for classifying emotions using customized decision tree algorithm," in *Proc. IEEE Southeastcon*. IEEE, 2012.

[20] P. Ekman, "Cross-cultural studies of facial expression," *Darwin and facial expression: A century of research in review*, vol. 169222, p. 1, 1973.

[21] L. Alam and M. Hoque, "A text-based chat system embodied with an expressive agent," *Advances in Human-Computer Interaction*, pp. 1–14.