IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Emotion Detection from Punjabi Text using Hybrid Support Vector Machine and Maximum Entropy Algorithm

Er. Ubeeka Jain[1], Amandeep Sandu[2]

Department of Computer Science & Engg, RIEIT RailMajra[1, 2]

Abstract: Emotion detection is the approach to detect the human emotion from different ways of images, text, videos and audios etc. We are focusing on the emotion detection from Punjabi text language. Vast amount of work has been done for the English language. In spite of lack of resources for Indian languages, work has been done for Telugu, Bengali and Hindi language. Punjab is one of Indian states and Punjabi is its official language. Punjabi is under resourced language. In this paper, we proposed a hybrid research approach for the emotion detection of the Punjabi text. The hybridization involves the concept of Support Vector machine and Maximum Entropy Algorithm. The dataset is considered from the Punjabi websites, news paper and various Punjabi blogs. The combined form of dataset is considered for the emotion detection. In this paper, we have considered six classes of emotions as Joy, Sadness, Fear, Surprise, Disgust and Anger. The goal of this research paper is to classify the considered dataset into the form of these six emotions.

Keywords: Emotion Detection, Punjabi dataset, Support Vector Machine, Maximum Entropy Algorithm, Emotions.

I. INTRODUCTION

Today, Punjabi is widely used and well spoken language in the various parts of world. This language has 100+ million speakers and wide coverage area across the web. But the thing, which is scarce, is the resources and tools to do successive research in this language [1]. In his research work, we have developed a dataset for emotion detection in Punjabi language. The dataset is developed from the Punjabi websites and blogs.

Nowadays in the web there is a large amount of textual information. It is interesting to extract emotions for different goals like those of business. For example, in luxury goods, the emotional aspects as brand, uniqueness and prestige for purchasing decisions, are more important than rational aspects such as technical, functional or price. In this case customer is happy to buy a product even with high prices. Emotional Marketing aims to stimulate emotions in customer for tying him to brand and so increase the sell of product/service. Nowadays it isn't the product to be sold, since for each category there is a wide choice, but the focus is the relationship that the consumer establishes with the brand and with the emotions which the product communicates [2].

An emotion is a particular feeling that characterizes a state of mind, such as joy, anger, love, fear and so on. Automatic emotion detection from text has attracted growing attention due to its potentially useful applications. Emotion recognition from text has many applications. Consider for example an employee sending a harsh email to his colleague or superior. A tool that can analyze the email for emotions and alert the employee about its harshness before sending it comes in very handy to protect the employee's state. Consider also an emotion-based search engine that ranks documents according to the emotion requested by the user. Such an engine could prove to be very beneficial to users in a certain emotional state and can improve the effectiveness of the information retrieval process. Other useful tools that can benefit from emotion recognition from text include recommender systems that aim to personalize recommendations based on the user's emotions [3] [4].

Recognizing user's emotions is a major challenge for both humans and machines. On one hand, people may not be able to recognize or state their own emotions at certain times. On the other hand, machines need to have accurate ground truth for emotion modeling, and also require advanced machine learning algorithms for developing the emotion models [5].

In this research paper, we have proposed a hybrid approach of Support vector machine and Maximum entropy algorithm for the classification of emotions from Punjabi text form. The dataset is considered from the Punjabi online blogs. The six classes of emotions are taken as Joy, Surprise, Sadness, Fear, Disgust and Anger.

The rest of the paper is organized in the following manner: Section II describe the basic concept involved in the hybridization concept. Section III explains the proposed concept. Section IV gives the result values and comparison with other concepts. Section V concludes the paper.

II. BASIC CONCEPTS

This section covers the concept of basics of Support Vector Machine and Maximum Entropy Algorithm. These two concepts are explained as below:

A. Support Vector Machine

Support Vector machine is a system that receive data as input during a training phase, build a model of the input and output a hypothesis function that can be used to

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

predict future data [6]. SVM is supervised learning model with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis. Support Vector Machines, instead of being based on heuristics or analogies with natural learning systems, are based on results from statistical learning theory [7]. Thus, theoretical guarantees can be made about their performance by placing an upper bound on the generalization error using the VC-dimension of the classes of the learned functions, essentially performing structural risk minimization. SVMs perform an implicit embedding of data into a high-dimensional feature space, where linear algebra and geometry may be used to separate data that is only separable with nonlinear rules in input space

B. Maximum Entropy Algorithm

The Principle of Maximum Entropy is based on the premise that when estimating the probability distribution, you should select that distribution which leaves you the largest remaining uncertainty (i.e., the maximum entropy) consistent with your constraints [8]. That way you have not introduced any additional assumptions or biases into your calculations. The motivating idea behind maximum entropy is that one should prefer the most uniform models that also satisfy any given constraints [9]. For example, consider a four-way text classification task where we are told only that on average 40% of documents with the word "professor" in them are in the faculty class. Intuitively, when given a document with "professor" in it, we would say it has a 40% chance of being a faculty document, and a 20% chance for each of the other three classes. If a document does not have "professor" we would guess the uniform class distribution, 25% each. This model is exactly the maximum entropy model that conforms to our known constraint. The probability of MEA can be calculated as:

$$P_{ME}(c|d,\lambda) = \frac{\exp[\sum_i \lambda_i f_i(c,w)]}{\sum_{c'} \exp[\sum_i \lambda_i f_i(c,w)]}$$

Where, c' is the class value, w is the word and λ is the weight vector.

In its most general formulation, maximum entropy can be used to estimate any probability distribution.

III.DATASET CONSIDERED

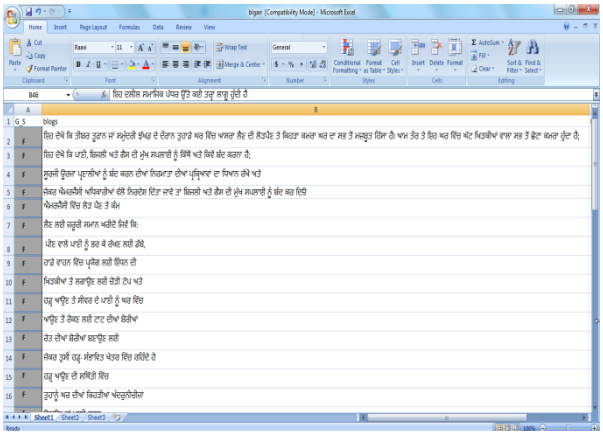

Figure 1: Punjabi language Dataset

WE have considered a dataset of Punjabi language from different Punjabi websites and blogs available online [10] [11] [12]. This dataset contains emotions in different form. On manual observations of the dataset, we have considered the 6 emotions that can be estimated from the considered dataset. A sample of this dataset is as shown in figure 1.

This dataset contains emotions in different form. Also the age and gender factor included for the emotion detection, because after the age of 70, emotions of surprise and joy decreases in most of the cases. And emotions of sadness are less seen to have in age less than 10.

IV. PROPOSED CONCEPT

In this section, hybrid concept of support vector machine and maximum entropy algorithm is defined. In this emotion detection, Support vector machine is considered to define the values in vector matrix form which can be calculated from linear regression model. These vector values in matrix form are used to decide the value of emotion by defining a threshold value. Further depending upon these values, emotion is categorized into the class of Joy, Surprise, Anger, Sadness, fear, Disgust.

Algorithm

Step 1: Consider the training dataset having data in Punjabi language text form.

Step 2: Set the counter to 1 and initialize the emotion extraction process from 1st sentence.

Step 3: Preprocess the data by applying the functions of stemmer and stop function.

3.1. Develop Array of Suffixes which are to be identified and removed to get a Root Word. A predefined list of root words is considered for the considered text database.

3.2. Boolean = Check word to be Stemmed Exists in the Dictionary.

3.3. If Boolean = true then no Stemming required.

Else

Root Word=get Root Word (Word to be Stemmed)

3.4. Check if Suffix exits in the Suffix Array developed in Step 3.1.

Replace Suffix (Word to be Stemmed Suffix)

Else

Go to next word.

3.5. Remove the words of stop function.

Step 4: Apply the concept of Support Vector Machine for classifying the sentences into vector values for linear regression model. The training dataset examples considered are as:

$$S = ((x_1, y_1), \ldots (x_i, y_i))$$

where the x_i are input data and the y_i are class labels, learning systems typically try to find a decision function of the form

$$h(x) = \mathrm{sgn}(\langle w.x \rangle + b)$$

(Where w is a vector of weights and b is called the bias) that yields a label $\in \{-1,1\}$ (for the basic case of binary classification) for a previously input x.

Step 5: The division of classes depends upon the value of

DOI 10.17148/IJARCCE.2015.41121

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

linear regression model. The vector values for the linear regression model can be calculated as:

$$Y_p = mX + b$$

where Yp is the predicted value of the dependent variable, m is the weight map of the previous input, and b is the bias value.

Step 6: Store these vector values of sentences, stop functions and stemmer function into matrix form.

Step 7: Apply the concept of Maximum Entropy Algorithm.

7.1. Consider the weight vector 'λ' to check the value as per the threshold value 't'. Weight vector decides the significance in the classification.

7.2. Decide the state by calculating the Maximum Entropy Probability function as defined below:

$$P_{ME}(c|d,\lambda) = \frac{exp[\sum_i \lambda_i f_i(c,w)]}{\sum_{c'} exp[\sum_i \lambda_i f_i(c,w)]}$$

Where, c' is the class value, w is the word and λ is the weight vector.

7.3. Define the classes of the emotions by declaring a threshold value 't' and categorize the classes in the form of Sadness, neutral and happiness.

7.4. If 'λ' is less than 't' (λ<t), then Emotions are more towards the state of Negative emotions.

7.5. If 'λ' is more than 't' (λ>t), then Emotions are more towards the state of Positive emotions.

7.6. If 'λ' is equal to't' (λ=t), then Emotions are more towards the state of neutral value.

Step 8. Probability function tells the further classification of the value after attaining the weight factor. More the probability value gives the more strength of the class.

Step 9: According the obtain values; put the emotion into their respective classes.

Step 10: Continue still all the dataset emotions are evaluated.

V. RESULTS & DISCUSSION

This section examines the various calculated parameters along with their results of the emotions.

A. Results

When hybrid algorithm is applied to the emotions of various Punjabi texts, then the output of text is described in the form of Joy, Sadness, Surprise, fear, Disgust and Anger. The calculated results of different emotions are shown in figure 2 as below:

More the value of the Emotions more will be the color spread of that emotion. Here, Red star symbol indicate the value of sadness, black cross sign indicate the value of Joy, Blue circle is for Disgust, Inverted triangle symbol of cyan color for surprise, pink triangular symbol for fear and small black dot star symbol for Anger.

Here, most of the emotions are representing their different value. The overlapping of two colors defines the emotion ambiguity, i.e. sentence is defining the emotions in two respects. For example, there can be various emotions to define happiness that either can be joy or surprise etc.

To define the accuracy of these emotions, we have calculated the lower bound, upper bound and average value of the each emotions with the percentage of error

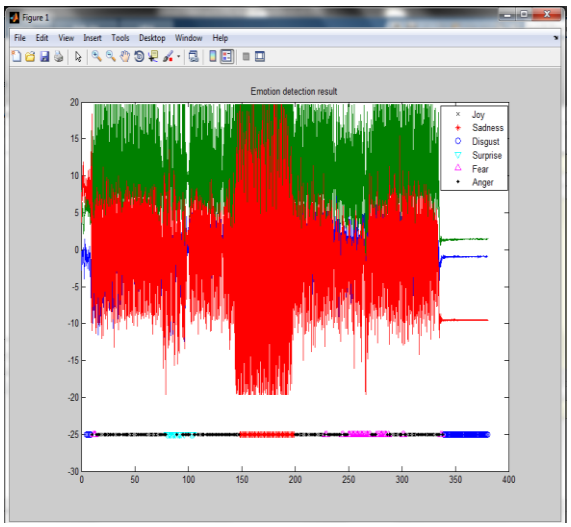

Figure 2: Detected Emotions from Dataset

matrix. The percentage of error matrix defined he probability of emotions that have more than one feature i.e. emotions that lie in more than one class. This can be shown in figure 3.

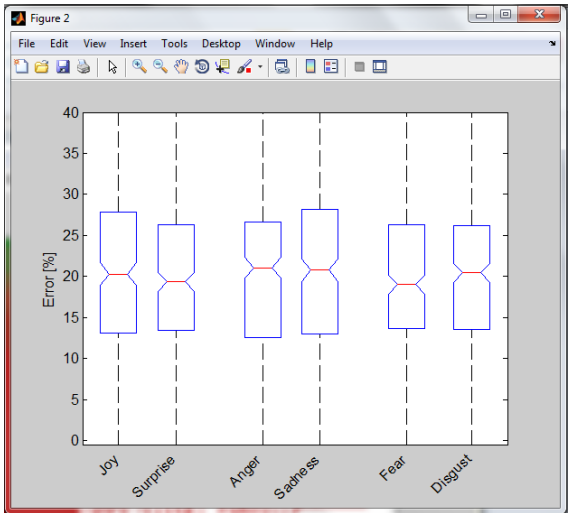

Figure 3: Probability distribution of emotions with respect to error matrix

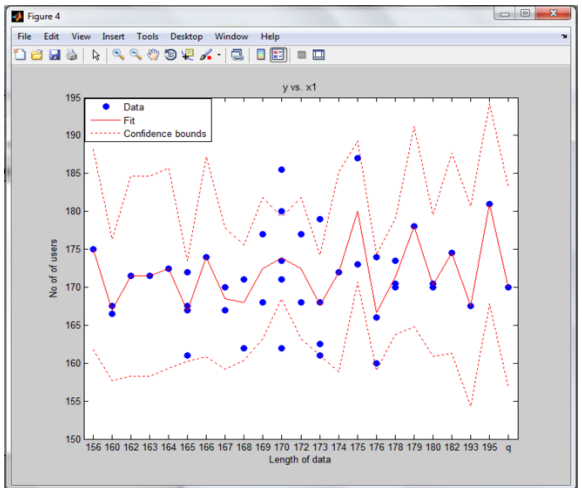

Figure 4: Fit bound for considered range of emotions

DOI 10.17148/IJARCCE.2015.41121

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Further, the overall value of the considered dataset is checked by considering the lower bound and upper bound of the Maximum Entropy Algorithm. These bounds define that whether the considered data lie in the emotional range or not.

In this figure 4, data is represented by blue dots values, lower and upper bounds represented by dot lines and linear graph line indicate the fit confidence level of the data. The data that is represented on the fit line, shows the emotions are perfectly matched. Emotion data above the fit line indicate data with more positive value and below line indicate data with more negative emotions.

From the proposed concept, the calculated estimate probability of the each emotion as described below as in figure 5:

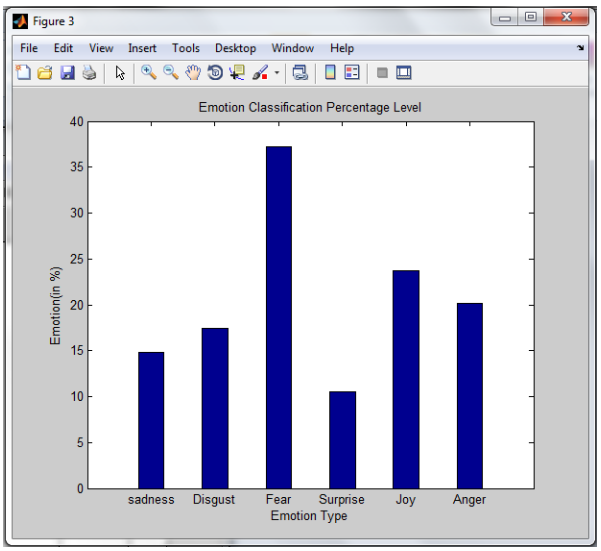

Figure 5: Estimate probability of each emotion

The exact value of these above emotions can be seen by the table 1 given below:

Table1: Calculated value of each emotion

Emotion	Percentage
Joy	23.755
Surprise	10.510
Sadness	14.858
Fear	37.213
Disgust	17.435
Anger	20.178

Further, Emotions are shown in the form of confusion matrix. A Confusion Matrix is expressed in tabular form with data in rows and columns classification accuracy of emotions is assessed. Confusion matrix describes a relationship between the absolutely classified emotions and erroneous emotions for each emotional class. The diagonal elements show the actual values of the emotions as shown in figure 6. Here. Total value for each emotion is considered as 1. The actual calculated value of these emotions for joy is 0.96,surprise is 0.69, anger is 0.66, sadness is 0.51, disgust is 0.80 and fear is 0.67. Rest values are described in figure 6.

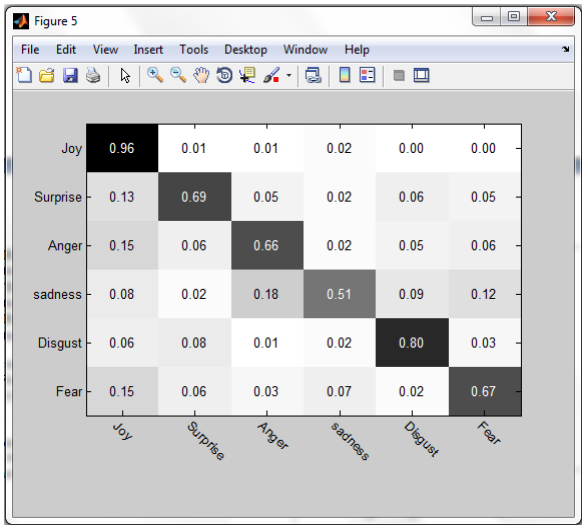

Figure 6: Confusion Matrix of Emotions

B. Evaluation parameters

There are various metrics used to evaluate different techniques. In our work we used the parameters performance of Precision, Recall and F-Measure.

To calculate the Precision, Recall and F-Measure, we have to understand the following terms as below:

- **TP (True Positive)**: is the number of positive detected emotions, which are classified as such.
- **FN (False Negative)**: is the number of negative detected emotions, which are not classified as such.
- **TN (True Negative)**: is the number of negative detected emotions, which are classified as positive.
- **FP (False Positive):**is the number of positive detected emotions, which are classified as negative.

We keep a record of some important measures which are TP, FN, TN, FP. From these we compute the measures Precision (p), and Recall (r), F-measure which are defined as follows:

(1). Precision

Precision denotes the probability that detected emotions from the dataset are truly detected. It is denoted by symbol **p**. It can be calculated as below:

$$p = \frac{TP}{TP + FP}$$

(2). Recall

Recall shows the probability that the emotions are actually detected. It is denoted by symbol **r.** It can be calculated as below:

$$r = \frac{TP}{TP + FN}$$

(3). F-Measure:F-measure is the combination between recall and precision. The Precision and Recall are further used to calculate the value of F- measure. F measure can be calculated asbelow:

$$F = \frac{2pr}{p + r}$$

The calculated values of Precision, Recall and F-Measure can be shown in table 2 given below:

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Table 2: Results of Emotion detection Parameters

Emotion	Precision	Recall	F-Measure
Joy	61.538	88.888	72.727
Surprise	71.428	100.00	83.333
Sadness	83.333	21.739	34.482
Fear	57.142	80.00	66.666
Disgust	60.00	50.00	54.545
Anger	62.500	100.00	76.923

These values of emotion dataset can be shown as in the diagrammatic graph below:

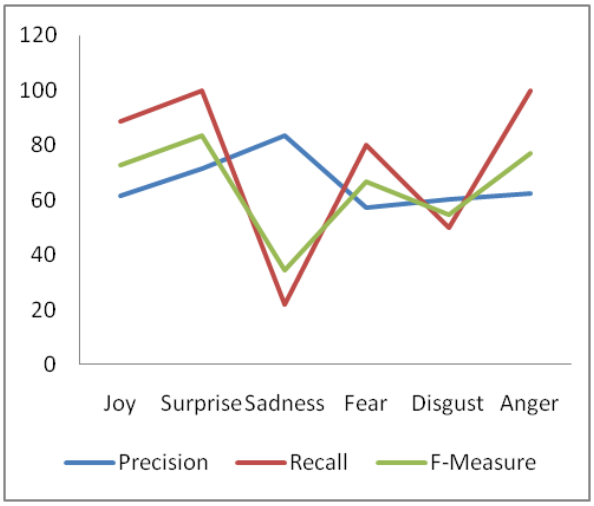

Figure 7: Parametric evaluation of Emotions

VI. CONCLUSIONS

In this paper, we proposed a hybrid research approach of Support Vector machine and Maximum Entropy Algorithm for the emotion detection of the Punjabi text. The dataset is considered from the Punjabi websites, news paper and various Punjabi blogs. The resources used to generate the Punjabi dataset had to be done from scratch because no work had been done in the area. The combined form of dataset is considered for the emotion detection. The main focus of the research work is to classify the emotions into the considered classes of Joy, Anger, Surprise, Sadness, Fear and Disgust. For the evaluation of the results, we have calculated the values of Precision, Recall and F-Measure for each respective emotion which can be shown by table 2 and figure 7. Also the calculated value of each emotion is shown as in figure 5 and table 1. From our proposed concept and considered dataset, we have calculated the values of emotions like joy, surprise, sadness, fear, disgust & anger are 23.755, 10.510, 14.858, 37.213, 17.435 & 20.178 respectively.

REFERENCES

[1]. Lehal, Gurpreet Singh. "A Survey of the State of the Art in Punjabi Language Processing." Language In India 9, no. 10 (2009): 9-23.

[2]. Shivhare, Shiv Naresh, and Saritha Khethawat. "Emotion detection from text."arXiv preprint arXiv:1205.4944 (2012).

[3]. Binali, Haji, Chen Wu, and Vidyasagar Potdar. "Computational approaches for emotion detection in text." In Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on, pp. 172-177. IEEE, 2010.

[4]. Munezero, Myriam, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text." Affective Computing, IEEE Transactions on 5, no. 2 (2014): 101-111.

[5]. Bijalwan, Vishwanath, Pinki Kumari, Jordan Pascual, and Vijay Bhaskar Semwal. "Machine learning approach for text and document mining." arXiv preprint arXiv:1406.1580 (2014).

[6]. Rebentrost, Patrick, Masoud Mohseni, and Seth Lloyd. "Quantum support vector machine for big data classification." Physical review letters 113, no. 13 (2014): 130503.

[7]. Agrawal, Ankit, and Aijun An. "Unsupervised emotion detection from text using semantic and syntactic relations." In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on, vol. 1, pp. 346-353. IEEE, 2012.

[8]. Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra. "A maximum entropy approach to natural language processing." Computational linguistics 22, no. 1 (1996): 39-71.

[9]. Wicentowski, Richard, and Matthew R. Sydes. "emotion Detection in suicide notes using Maximum Entropy Classification." Biomedical informatics insights5, no. Suppl 1 (2012): 51.

[10]. http://www.punjabitribuneonline.com

[11]. http://www.parchanve.wordpress.com

[12]. http://www.dailypunjabtimes.com

DOI 10.17148/IJARCCE.2015.41121