ISSN 2278-3091
Volume 8, No.5, September - October 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse54852019.pdf
https://doi.org/10.30534/ijatcse/2019/54852019

Dynamic Behavior Extraction from Social Interactions Using Machine Learning and Study of Over Fitting Problem

Sushil Kumar Trisal [1], Ajay Kaul [2]
[1] Department Of Computer Science And Engineering, Shri Mata Vaishno Devi University, Katra, Jammu And Kashmir, India. Email: sushil.trisal@ smvdu.ac.in
[2] Department Of Computer Science And Engineering, Shri Mata Vaishno Devi University, Katra, Jammu And Kashmir, India. Email: ajay.kaul@ smvdu.ac.in

ABSTRACT

People in this world interact with each other directly (face-to-face) or indirectly (by sending text messages). As humans are linked with emotions themselves it is easy for them to detect emotions but same is not true with computers. But for a computer detecting an emotion will be a difficult job to perform. Emotion is cognitive process where machine learning techniques are used to detect it. The main goal of the machine learning algorithms is to build computer system that can adopt and learn from their experience or by examples. This paper presents a systematic performance analysis of four classification algorithms for the extraction of human emotions and does also over fitting problem analysis. Our main attempt in this paper is to detect emotions from social interactions on twitter social network. Four machine learning algorithm are used for text classification under seven emotional classes. The work is carried on two data bases one data base is build by collecting live tweets from twitter social network site and second ISEAR (International Survey on Emotion Antecedents and Reactions) database. WEKA interface is used for the implementation of our work which shows above 85% accuracy in identifying the emotion classes. The ranking and standard deviation functionalities provided by the WEKA experimenter helps to determine the effectiveness of a classifier model.

Key words: Machine Learning; Human Emotions; WEKA (Waikato Environment for Knowledge Analysis); Affective computing; International Survey on Emotion Antecedents and Reactions (ISEAR).

1. INTRODUCTION

The Emotion is a mental state that arises spontaneously rather than through conscious effort and is often accompanied by physiological changes; a feeling: the emotions of joy, sorrow, reverence, hate, and love. Crying, blushing, laughing or a variety of facial expressions are evident and outward are known as physical form of emotion. The emotions can be specified through facial expressions, textual information, or speech. The widespread form of communication on web is in the form of text. The emotion inferred from this textual data is useful in many areas such as sentiment analysis, text to speech generation, and better computer interaction system. Emotions may be expressed by a single word or a group of words. In computational linguistics, the automatic emotion detection from texts is becoming increasingly important from an applicative point of view. Emotion classification allows us to identify the feelings of individuals toward specific events. The most natural way for automatic emotion recognition of the user is to detect his emotional state from the text that he entered in a blog, an online chat site, or in another form of text. Emotion process is very complex; it consists of various components which include cognitive behavior, changes in body actions, feeling and thoughts. These components interact with each other to give rise new models of emotions, but there is not a single universally acceptable model at present. The first step in modeling any phenomenon is data collection in machine learning process. Machine learning or cognition is the process of acquiring and understanding knowledge through our thoughts, experience, and senses. Whenever you see or hear something new, you go through series of cognitive processes, which are the processes that result in learning. Learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. There are different learning styles or learning models that an algorithm can adopt or suit with the type of the problem. A brief definition of these learning methods is given below.

Supervised Learning: In this type of learning the input data which is called training data is labeled. The examples of such labeled data are like spam/not-spam or a price of stock exchange at a time. Through a training process model is prepared and the predictions are corrected when they are wrong. The training process stops when the desired level of accuracy is reached.

Unsupervised Learning: In this type of learning the result is not known because input data is not labeled. Structures in the input data are used to build the model in unsupervised learning. Association rule learning and clustering examples of this learning. Input data is not labeled and does not have a known result. Apriori algorithm and k-means are example algorithms.

Semi-Supervised Learning: The training data in this type of learning is a mixture of labeled and unlabeled data. The model learns from the structures to organize both the data as well as make predictions. Classification and regression are the examples of the Semi-Supervised Learning.

Reinforcement Learning: In this case of learning labeled data is provided as stimulus to a model which feeds as input - environment to which the model must respond and react. As in supervised learning feedback is provided not for teaching process, but as punishments and rewards in the environment. Systems and robot control are example problems of

Reinforcement Learning. Q-learning and temporal difference learning is example algorithms.

In this way of organizing machine learning algorithm is useful for selecting input data. It gives us vision to think about model preparation, data, and process to select one that is associated with the most appropriate for your problem in order to get best and true results.

In this paper, our approach is to classify the emotions in text. It is focusing on identifying seven different classes of emotion such as disgust, anger, guilt, fear, joy, shame and sadness.

The remainder of this paper is organized as follows: Section II introduces back ground research in this area. The III section defines our motivation and contribution. The IV section presents our proposed system model. Section V is about the experiments and detailed result analysis, which is followed by the conclusion and Future work. Last section is the references.

2. BACK GROUND RESEARCH

Researchers have been fascinated by emotions which are very evident from the research body working in the fields of linguistics, psychology, social sciences and communication. Consequently, along several directions the research on emotions has been pursued. Various methods and models have been developed for the emotion detection from the text. As the number of users are increasing day by day on social networks it is being more open for the researchers as large volume of information is available. To theorize emotions great contribution is from the French philosopher René Descartes' Treatise, Les passions de l'âme (Passions of the Soul), published in 1649. The author contributes to a long tradition of theorizing "the passion". The passions were experiences now commonly called emotions in the modern period. Tomkins, in 1962 was a psychologist and personality theorist who developed both affect theory and script theory. R.W.Picard [1] discussed the key issues in affective computing, computing that relates to arise from, or influences emotions. In the literature several works have been reported Izard, 1977; Robert Plutchik, 1980 "The Nature of Emotions" [2] agrees that the study of emotion is one of the most confused (and still open) chapters in the history of psychology. Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Basic emotions which are universally accepted are given by Ekman, 1992 [3] based on facial expressions. According to him six basic emotions have been defined as fear, anger, disgust, sadness, surprise and happiness. Liu et al., 2003; Alm et al., 2005; and Neviarouskaya et al., 2007a,b. Two main aspects positive and negative effect, was identified by the Circumplex Theory of Affect (Watson and Tellegen, 1985) [4] which range from high to low. By analyzing 590 English words Johnson-Laird and Oatley (1989) [5] have deduced basic emotions by which describe emotion. Osgood's theory of Semantic Differentiation (Osgood et al., 1957) assigns three dimensions to emotive meanings to words along three directions of emotional classes. Implicit and explicit emotion of the words was define by (Clore et al., 1987). The emotion classification was defined on the contex of the words in the emotional class. Direct affective words and indirect affective words explicit and implicit categories was defined by Strapparava and Valitutti (2006) [6]. In this context of differentiation emotion reorganization was done by the key word classification in the emotion sentence. Kaveh Bakhtiyari & Hafizah Husain [7] discusses a fuzzy model for multi-level human emotions recognition by computer systems through keyboard keystrokes, mouse and touch screen interactions. Jianhua Tao et al [8] generated emotion estimation net (ESiN) that combined the content words and emotion functional words (EFWs) to estimate the final emotion output. They used the text from a spontaneous speech corpus and obtained relatively good results.

Yang et al [9] proposed the emotion classification of web blog corpora using Support Vector Machine (SVM) and Conditional Random Field (CRF) machine learning techniques.

A bag of words approach to emotion classification was introduced by Danisman et al [10]. They considered the ISEAR dataset and tested various classifiers including SVMs, Naïve Bayes and Vector Space Model (VSM), and found that VSMs gave most promising results.

Automatic emotion detection method based on Knowledge and Corpus was proposed by Strapparava et al [11]. They considered emotion analysis of news headlines, and proposed five systems for emotion extraction: WIN-AFFECT presence, LSA emotion synset, Latent Semantic Analysis (LSA) single word, LSA all emotion words, NB trained on blogs. They found that the Latent Semantic Analysis (LSA) system using all the emotion words per-formed well.

Dipankar et al [12] proposed a mechanism for sentence level emotion identification. They used the Conditional Random Field (CRF) classifier for the classification of the words into one neutral tag and the six emotion tags.

Bellegarda et al [13] described a method for emotion analysis based on the principles of latent semantics, using two techniques which are latent affective folding and embedding.

Paul N. Bennett [14] introduced a probabilistic method for combining classifiers that considers the context sensitive reliabilities of contributing. It harnesses reliability indicators.

2.1 Conclusion Literature Survey

In the literature several works have been reported on the emotion study under text, expressions using different approaches like text analysis with mathematical algorithms, emotion recognition methods using supervised machine learning algorithms [15], keystroke analysis and mouse movement [16], ECG and EEG pattern analysis [17] etc. As we know in text expression there are no facial expressions, gestures and like in image processing. Fuzzy based system given by Caryl Charlene et al [18] identifies performing and non-performing employees as organizations. Face Recognition System for multi-view vision system proposed by Jay Robert and B. Del Rosario [19]. The classifications of the emotions totally depend upon the creative words in the text communication. For this reason the machine learning algorithms are used in this work which is best suitable for the prediction and classification analysis for detection of emotions using text as input. The motivation for this work has also come from the recent increasing interest in the field of emotion analysis due to the increased online communication on social networks.

3. MOTIVATION AND CONTRIBUTION

Keeping trust, privacy and security issues in consideration on social networks it is very interesting to detect human emotions which are the main motivation factor. The second factor is that millions of users are integrated in these sites and are shearing a lot of information. Hence a large volume of information is available on social sites to detect the dynamic human behavior like noting behavior in upcoming elections, product taste, and growth of business and create new business and security of related information exchange.

Contribution of this work presents a systematic performance analysis of four classification algorithms used in Machine learning for the extraction of human emotions and does also over fitting problem analysis. Main attempt in this paper is to identify emotions which have been classified under seven emotional classes by text classification process using machine learning techniques. Machine learning usually refers to the changes in system that performs tasks associated with artificial intelligence (AI). Such tasks involve recognition, diagnosis, planning, robot control, prediction, etc. Four machine learning algorithms used for text classification in this paper are Naïve Bayes, J48, K Nearest Neighbors (KNN) and Support Vector Machine (SVM). Our approach shows above 85% accuracy in identifying the emotion classes. In the second part we have combined the algorithms Navie Bayes, J48 and KNN, Support Vector Machine for analyzing the over fitting problem. Overfitting problem is defined as when an algorithm it is more accurate in fitting known data and less accurate in predicting new data. The work is carried on two data bases one data base is built by collecting live tweets from twitter social network site and second ISEAR database using WEKA interface.

4. PROPOSED SYSTEM MODEL

The proposed system model is shown in figure 1. Our main task in this work is to detect user emotion by two ways, first four classifier algorithms namely Naïve Bayes, J48, K Nearest Neighbors (KNN) and Support Vector Machine (SVM) are used for emotion detection their performance is evaluated [20]. This work is carried on seven basic emotional classes – joy, anger, sadness, disgust, guilt, fear, and shame as per the basic emotional categories of ISEAR database. Application program Interface (API) was developed for collecting live tweets form twitter. Another important dimension of this research work is that it combines classifier algorithms Navie Bayes and J48, KNN and Support Vector Machine to study the over fitting problem. When model begins to "memorize" training data rather than "learning" Overfitting occurs. In order to detect emotions from text more effectively we implement the classifiers together by combination. Our work in this research concentrates on this dimension and is a potential field to work on.

4.1 Text Classification Process

The text classification process consists of the following steps

4.1.1 Documents Collection

In this step different types of documents with different formats are collected for the building of dataset. The different types of formats are like pdf, html, doc, etc. the two datasets in this work are from two databases one from ISEAR and second from twitter.

4.1.2 Pre-Processing

In pre-processing step the text document is formatted in to a word format document which is the first step of pre-processing. There are great number of features in this document which are processed by taking the following steps.

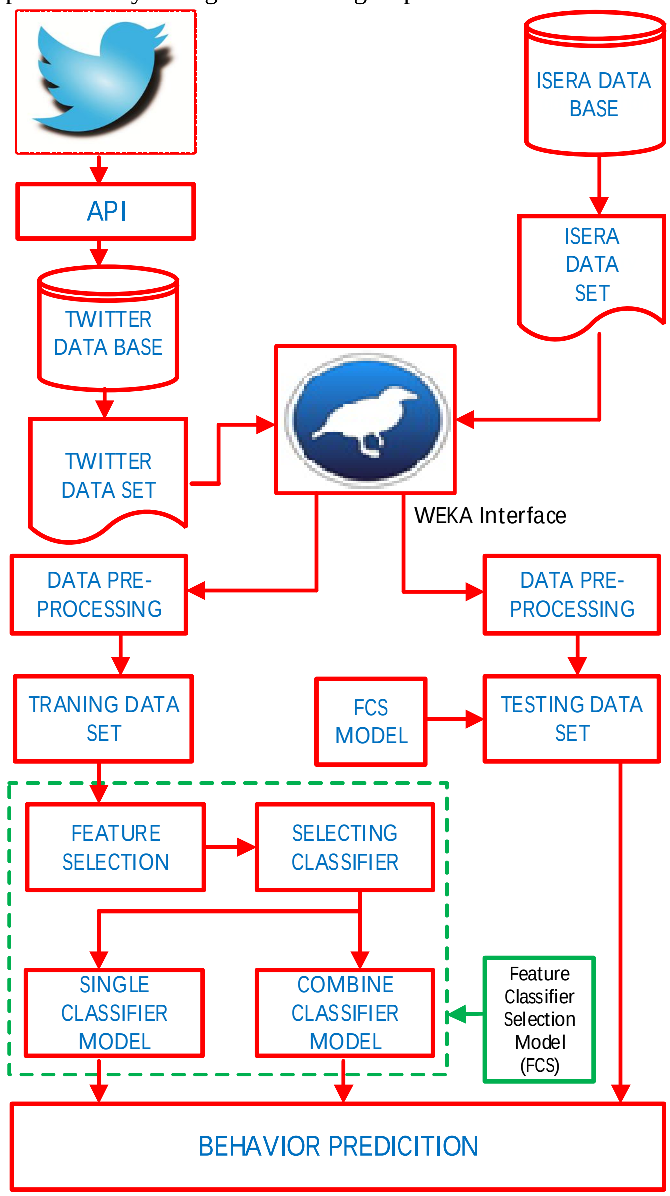

Figure 1: Shows proposed system model

The goal of the tokenization is the. But still there are some problems that have been left, for e.g., the removal of punctuation marks as well as other characters like brackets, hyphens, etc. The main use of tokenization is identification of meaningful keywords.

4.1.4 Stemming

Reduction of words in to its root or base word is called stemming. For e.g., many words in English can be reduced to is steam word like, liking, likely, unlike belong to like. By removing the "s" names can be transformed in to their steam word.

4.1.5 Stop word Elimination

Words that do not help in predication like articles, prepositions and nouns are considered as stop words. Any group of words or phrases can be chosen for this purpose. This step is very important as it increases the classification performance. Examples of stop words are “a”, “is”, “you”, “an”.

4.1.6 Indexing

This technique is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector The Perhaps most commonly used document representation is called vector space model.

4.1.7 Feature Selection

After pre-processing and indexing the important step of text classification, is feature selection to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier by detecting the irrelevant for classification of emotions. The main idea of Feature Selection (FS) is to select subset of features from the original documents. The feature selection procedure has a number of advantages, as it reduces the size of the dataset. Increase the efficiency of the emotion classifier. It reduces the computational complexity of the algorithm for text categorization and the search space is shrinked. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Another benefit of feature selection is its tendency to reduce over fitting.

4.2 Classification Methods

4.2.1 J48 Decision Trees

Decision tree algorithm is a simple yet widely used technique for classification. Decision tree has a tree structure similar to flow chart. This tree structure is consisting of three types of nodes.

a) Nodes that does not having any incoming edges but has zero or more than one outgoing edges is called a root node.

b) Nodes that are having one incoming edge and two or more outgoing edges are called internal nodes. Internal nodes denote test on all attributes in the tree structure.

c) Nodes that have one incoming edge and have zero out going edge are called leaf node in the tree structure. Class labels are denoted by leaf nodes

Usage of the decision tree is divided into two phases.

a) Tree Building
b) Tree Validating

Initially all the training records are at the root. The input to this algorithm consists of the training records E and the attribute set F. The algorithm works by recursively selecting the best attribute to split the data and expanding the leaf nodes until stopping criterion is met.

4.2.2 Naive Bayes Classifier

This is a very simple probabilistic algorithm used for classification of the emotional classes. The algorithm works by counting the number of frequency and combinations of the key words in the given sentence or a data set. Bayes theorem is used by Naive Bayes classification algorithm which assumes all attributes independent to the given class variable in the data set. All assumptions rarely holds true in real world applications but still the algorithm is having very exciting applications in the real world and performs better in various supervised classification problems.

As Bayes’ theorem is used by Naive Bayesian classifier is the theorem of total probability [6]. The probability that a document d with vector x = < x1... xn> which belongs to a hypothesis H1 is

$$P\left(\frac{H1}{Xi}\right) = P\left(\frac{Xi}{H1}\right) * P(H1) / P(\frac{Xi}{H1} * P(h1) + P\left(\frac{Xi}{H2}\right) * P(H2)$$

Here, Posterior probability is P (H1|xi), while is the prior probability is P (H1) which is associated with hypothesis H1.
For total number of “n” different hypotheses, we have

$$P(xi) = \sum_{j=1}^{n} P\left(\frac{xi}{hj}\right) * P(hj)$$

Therefore, we have

$$P\left(\frac{H1}{xi}\right) = (P\left(\frac{xi}{H1}\right) * \frac{P(H1)}{P(Xi)})$$

In comparison with other algorithms Naïve Bayes classifier algorithm works on works on intuitive concept which increases its efficiency for the detection of the emotions. It is seen in various applications that comparatively Naïve Bayes outperforms many other complex classification algorithms as it makes use of that variables in data sample individually, independent of each other which in turn increases the computational complexity of the algorithm.

4.2.3 K-Nearest Neighbors (KNN)

This algorithm uses the local neighborhood to obtain a prediction. The K memorized examples more similar to the one that is being classified are retrieved. A distance function is needed to compare the examples similarity. In K-Nearest-Neighbor (KNN) algorithm distance is measured between query scenario and a set of scenarios in the data sample.
Distance function d(x,y) [6] is used to compute the distance between two scenarios.
x, y are scenarios including N features, such that

$$X = \{X1, \dots\dots\dots\dots, Xn\}$$
$$Y = \{Y1, \dots\dots\dots\dots, Yn\}$$

Two distance functions are:
Absolute distance measuring:

$$dA(X,Y) = \sum_{i=1}^{n} \left| Xi - yi \right|$$

Euclidean distance measuring:

$$dE(X,Y) = \sum_{i=1}^{n} \sqrt{Xi^2} - Yi^2$$

To reduce the effect of noise in the classification we generally choose the larger value of the ‘K’. the best choice of *k* depends upon the data sample itself. Various heuristic techniques are used for choosing the value of ‘K’. When ‘K=1’is called the special case of KNN.

4.2.4 Support Vector Machine (SVM)

In case of Support Vector Machines (SVM) classes are separated in a multidimensional space by constructing the hyper-plan. The SVM is based on decision plans in the given data space by constructing the decision boundaries. The

decision plan is defined set of objects in the decision boundaries in Support Vector Machine (SVM). Both regression and classification tasks are supported by SVM and can handle both categorical variables and continues data iterations.
An iterative training algorithm is used to construct the optimal hyper-plane in SVM. By this technique the error function is minimized. On the bases of error function SVM classification algorithm are classified under four groups which are defined below.

Classification SVM Type 1 (also known as C-SVM classification) for this type of SVM, training involves the minimization of the error function [22]:

$$\frac{1}{2} * W^T * W + C \sum_{i=1}^{N} \varepsilon i$$

Classification SVM Type 2 in contrast to Classification SVM Type 1 minimizes the error function model minimizes the error function [22]:

$$\frac{1}{2} * W^T * w - \nu\rho + \frac{1}{N} * \sum_{i=1}^{N} \varepsilon i$$

Regression SVM Type 1 or SVM regression. For this type of SVM the error function is [19]:

$$\frac{1}{2} * W^T * W + C * \sum_{i=1}^{N} \varepsilon i \; + C * \sum_{i=1}^{N} \varepsilon i$$

Regression SVM Type 2 or-SVM regression. For this SVM model, the error function is given by [22]:

$$\frac{1}{2} * WT * W - C[V\varepsilon + \frac{1}{N} * \sum_{i=1}^{n} (\varepsilon i + \varepsilon i)]$$

5 EXPERIMENTAL SETUP AND RESULT ANALYSIS

5.1 Data sets

Two data bases are used in this work. One data base is collected by streaming live tweets from twitter social network. An Application Program Interface is developed using Natural Language Tool Kit (NLTK). Approximately on lakh (1,00,000) tweets with user ID, Twitter Time, Retweet Count etc. where streamed. Five thousand (5000) tweets where used in our experiment for emotion detection purposes. The sample example tweets are shown in table 1.
ISEAR (International Survey on Emotion Antecedents and Reactions) is the second data base used. Data was collected for the ISEAR project by a group of psychologists directed by Klaus R. Scherer and Harald Wallbott. Experiences of a large number of people were collected for building the database in which situations where reported which include all the seven major emotion classes which are: anger, joy, sadness, fear, disgust, guilt and shame. The sample examples of ISEAR database are shown in table 2.

5.2 Experimental Setup

To analyze text, a standard dataset is needed upon which our classification algorithm is applied. For this purpose, we used International Survey on Emotion Antecedents and Reactions (ISEAR) dataset for text analysis (Scherer and Wallbott 1994). The final dataset contains more than 7500 emotion sentences collected for the emotion analysis. The dataset also includes the numeric values of the emotion sentences. The ISEAR dataset is used for conducting the experiment. The emotion tagged sentences in this dataset are first normalized by pre-processing. Then the sentences are tokenized for character n-gram extraction by using the tokenizer. Then the words are stemmed to their root form by stemming technique. The feature selection process is carried by TF-IDF ranking technique. The term TF is the Term Frequency and IDF means the Inverse Document Frequency. TF-IDF is computed as:

$$TF(W) = (Number\ of\ times\ term\ 'W'\ appears\ in\ a\ document) / (Total\ number\ of\ terms\ in\ a\ document\ )$$
$$IDF(W) = Log(Total\ number\ of\ documents) / (Number\ of\ documents\ with\ term\ W'\ in\ it)$$

The whole ISEAR dataset is used for training the machine learning algorithms to build the training classification model.

TABLE 1: SAMPLE TWEETS FROM TWITTER

Emotion	Example Sentence
Joy	“@riancurtis: I am, friend and I love you.”
Fear	“I think my arms are sore from tannins."
Anger	“@angry_barista I baked you a cake but I ated it "
Sadness	“@Starrbby too bad I won't be around I lost my job and can't even pay my phone bill lmao aw shucks”
Disgust	“ I hate when my baby is sick.”
Shame	“I feel bad for doing it the creepiest. Feeling.”
Guilt	“I still can not find my keys.”

This training classification model is implemented on the testing data set which are the tweets collected from twitter social network.
WEKA 3.7.12 [22], [24], [25] is an open source interface which is used for the classification. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. The Weka suite is the collection of visualization tools, graphical user interfaces and is used for predictive modeling. Several standard data mining tasks are supported by weka software. Other tasks supported are more specifically, data clustering, pre-processing, classification, visualization, regression. The data accepted in Weka is assumed as a single file or relation and each data is described by a fixed number of attributes (numeric or nominal). Datasets in WEKA accepts the data in .arff format that is attribute relation file format. Though it can accept data in CSV

TABLE 2: ISEAR DATABASE SAMPLE EMOTION SENTENCES

Emotion	Example Sentence
Joy	"When I pass an examination which I did not think I did well.”
Fear	“When for the first time I realized the meaning of death."
Anger	"Friends who torture animals!”
Sadness	“Breaking up with a girl."
Disgust	“When someone makes advances that one does not want."
Shame	“When I boxed my child's ears."
Guilt	"When I was fined for speeding.”

format also and can be converted into ARFF format. Ten fold cross validation approach is used for training the model. This model is trained on all folds except one that is left out for testing the algorithm. This process is repeated so that each fold gets an opportunity at being left out and acting for test data set. Lastly, the performance measures are averaged across all folds to estimate the capability of the algorithm on the test data

5.3 Results

5.3.1 Cost Insensitive Analysis

The test set is used to evaluate the performance of the classifiers. The dataset is classified into seven emotion categories: joy, fear, anger, disgust, guilt, sadness, and shame. Some standard terms used for measure under this analysis are:

i) TP (True positive): Measures the proportion of positives that are correctly identified.

ii) FP (False positive): It is the error in data reporting in which a test results improperly indicates presence of a condition which in reality it is not.

iii) Precision: It is the fraction of retrieved instances that are relevant.

$$Precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

iv) Recall: It is the fraction of relevant instances that are retrieved.

$$Recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

v) F-measure: A measure that combines precision and recall is the harmonic mean of precision and recall the traditional F-measure of balanced F-score.

$$F = \frac{2(Precision * Recall)}{(Precision + Recall)}$$

vi) MCC: Matthews correlation co-efficient it takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of different size. It returns a value between -1 and +1.

vii) ROC: Receiver operating characteristics or curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

viii) PRC Area: It represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

Our experiments show different accuracy in detecting the emotion classes from social media interactions for the machine learning algorithms. Class wise detail accuracy are shown in table 3 for J48 algorithm at 92.8571%, table 4 for Naive Bayes algorithm at 85.7143%, table 5 for K-Nearest Neighbors algorithm at 81.6326% and table 6 for Support Vector Machine algorithm at 98.5714%. The Support Vector Machine (SVM) algorithm shows the best accuracy at 98.5714% for the test data set as they can significantly reduce the need for labeled training instances. For over fitting problem analysis we combine the algorithms Naive Bayes and J48 and K-Nearest Neighbors and Support Vector Machine by voting technique. Table 7 shows the detail accuracy for Naive Bayes and J48 algorithms at 75.000% and Table 8 shows the detail accuracy for K-Nearest Neighbors and Support Vector Machine algorithms at 82.1429%. This analysis shows that the algorithms are fitting i.e., they are high accurate on training data but is less accurate in predicting new data sets [31].

Table 3: The simulation results for test set using J48 decision tree algorithm.

Re-evaluation on test set's
Correctly Classified Instances 4640 92.8571 %
Incorrectly Classified Instances 360 7.1429 %
Coverage of cases (0.95 level) 100 %
Total Number of Instances 5000
Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Joy
0.750	0.042	0.750	0.750	0.750	0.708	0.969	0.861	Fear
1.000	0.042	0.800	1.000	0.889	0.876	0.979	0.800	Anger
0.750	0.000	1.000	0.750	0.857	0.849	1.000	1.000	Sadness
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Disgust
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Shame
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Guilt
Weighted Avg.								
0.929	0.012	0.936	0.929	0.928	0.919	0.993	0.952	

5.3.2Cost Sensitive Analysis

This analysis of the classifier is used when the data set is highly skewed. The cost insensitive techniques precision, recall and F-Measure will not yield good results for such data sets. Receiver Operating Characteristics (ROC) graph or curve is used to illustrate the performance of the classifier model. The curve is created by plotting (at various threshold settings) the true positive rate (TPR) against the false positive rate (FPR) shown in Figure 2

Table 4: The simulation results for test set using Naive Bayes Classifier algorithm.

Correctly Classified Instances 4285 85.7143 %
Incorrectly Classified Instances 715 14.2857 %
Coverage of cases (0.95 level) 92.8571 %
Total Number of Instances 5000
Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.750	0.000	1.000	0.750	0.857	0.849	0.990	0.950	Joy
0.750	0.000	1.000	0.750	0.857	0.849	0.990	0.950	Fear
1.000	0.083	0.667	1.000	0.800	0.782	0.979	0.900	Anger
0.750	0.000	0.600	0.750	0.667	0.609	0.979	1.917	Sadness
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Disgust
0.750	0.000	1.000	0.750	0.857	0.849	0.906	0.827	Shame
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Guilt
Weighted Avg.								
0.857	0.024	0.895	0.857	0.863	0.848	0.978	0.935	

The main goal of ROC graph is to have the curve lot to right upper corner towards 1(one).

As shown in the graph the value of the SVM algorithm is more towards 1 as compared to the other algorithms. We can also find the AUC (Area Under Curve) which is .985 in case of the SVM algorithm higher than all other algorithms in comparison.

Table 5: The simulation results for test set using K-Nearest Neighbors algorithm.

Correctly Classified Instances	4080	81.6326 %
Incorrectly Classified Instances	920	18.3673 %
Root mean squared error	0.2415	
Coverage of cases (0.95 level)	100 %	
Total Number of Instances	5000	

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.167	0.500	1.000	0.667	0.645	0.940	0.643	Joy
0.714	0.000	1.000	0.714	0.833	0.826	1.000	1.000	Fear
0.857	0.119	0.545	1.000	0.667	0.619	0.974	0.798	Anger
0.714	0.048	0.714	0.714	0.714	0.667	0.913	0.491	Sadness
0.714	0.024	0.833	0.714	0.769	0.737	0.971	0.762	Disgust
0.286	0.024	0.667	0.286	0.400	0.382	0.923	0.608	Shame
0.429	0.000	1.000	1.429	0.600	0.626	0.932	0.664	Guilt
Weighted Avg.								
0.673	0.054	0.751	0.673	0.664	0.643	0.951	0.709	

Table 6: The simulation results for test set using Support Vector Machine algorithm.

Correctly Classified Instances	4925	98.5714 %
Incorrectly Classified Instances	75	1.4286 %
Relative absolute error	83.4921 %	
Root relative squared error	86.0941 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	74.0136 %	
Total Number of Instances	5000	

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Joy
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Fear
0.967	0.000	1.000	0.967	0.983	0.980	0.986	0.800	Anger
1.000	0.017	0.909	1.000	0.952	0.945	0.992	1.000	Sadness
0.967	0.000	1.000	0.967	0.983	0.980	0.980	0.986	Disgust
0.967	0.000	1.000	0.967	0.983	0.980	0.980	0.976	Shame
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Guilt
Weighted Avg.								
0.986	0.002	0.987	0.986	0.984	0.984	0.995	0.978	

Table 7: The simulation result for test set using Naive Bayes and J48 algorithm.

Correctly Classified Instances	4820	75 %
Incorrectly Classified Instances	180	25 %
Coverage of cases (0.95 level)	96.4286 %	
Total Number of Instances	5000	

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.500	0.000	0.500	0.500	0.500	0.417	0.938	0.708	Joy
0.750	0.042	0.750	0.750	0.750	0.708	0.927	0.841	Fear
1.000	0.042	0.800	1.000	0.889	0.876	0.990	0.950	Anger
0.500	0.0125	0.400	0.500	0.444	0.343	0.927	0.693	Sadness
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Disgust
0.750	0.000	1.000	0.750	0.857	0.849	1.000	1.000	Shame
0.750	0.000	1.000	0.750	0.857	0.849	0.979	0.917	Guilt
Weighted Avg.								
0.750	0.042	0.779	0.750	0.757	0.720	0.966	0.873	

Table 8: The simulation result for test set using KNN and Support Vector Machine algorithm.

Correctly Classified Instances	4105	82.1429 %
Incorrectly Classified Instances	895	17.8571 %
Coverage of cases (0.95 level)	96.4286 %	
Total Number of Instances	5000	

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.250	0.000	1.000	0.250	0.400	0.471	0.755	0.506	Joy
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Fear
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Anger
0.750	0.000	1.000	0.750	0.857	0.849	1.000	0.850	Sadness
1.000	0.208	0.444	1.000	0.615	0.593	1.000	1.000	Disgust
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Shame
0.750	0.000	1.000	0.750	1.000	0.849	0.964	0.861	Guilt
Weighted Avg.								
0.821	0.030	0.921	0.821	0.819	0.823	0.995	0.888	

5.3.3 Statistical Analysis

Evaluation of performance is compared using Mean absolute error, root mean squared error and kappa statistics. For good performance large test sets are used. For the classifiers as in case of the simulation result for test set using Support Vector Machine algorithm shown in table 6.

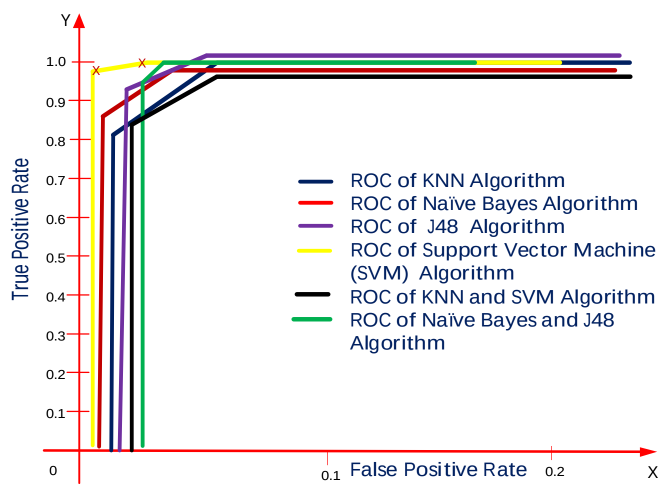

Figure 2: Shows ROC Graph.

The performance analysis in terms of kappa Statistics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are shown in table 9 and its graph is shown in figure 3.

Kappa Statistics (K)

Kappa is defined as normalized value of agreement for chance agreement

$$K = \{P(A) - P(E)\} \div (1 - P(E))$$

Where Percentage agreement = P (A), chance agreement=P (E).

When K =1 agreement is said to be perfect between the classifier and ground truth.

When K=0 that indicates there is a chance of agreement.

Mean Absolute Error

Quantity used to measure predictions of the eventual outcomes is the mean absolute error (MAE. The mean absolute error is given by $MAE = \frac{1}{n}\sum_{i=1}^{n} |fi - yi|$

$$MAE = \frac{1}{n} * \sum_{i=1}^{n} |ei|$$

The mean absolute error is an average of the absolute errors ei= |fi-yi| Where prediction = fi and true value = yi.

Root Mean Squared Error

The RMSE Ei of an individual program *i* is evaluated by the equation:

$$Ei = \sqrt{\frac{1}{n} * \sum_{j=1}^{n} (P(ij) - Tj/Tj)^2}$$

Where P (ij) = the value predicted by the individual program i = fitness case, Tj =the target value for fitness case j.

Table 9: Shows Kappa Statistic, MAE and RMSE values of different Machine Learning Algorithms.

Performance Analysis Classifier	Kappa Statistic	Mean Absolute Error	Root mean Squared Error
J48	0.9167	0.0264	0.1094
Naive Bayes Classifier	0.8333	0.0375	0.1732
K-Nearest Neighbors	0.619	0.1187	0.2415
Support Vector Machine	**0.9833**	**0.2045**	**0.3013**
Naive Bayes and J48	0.7083	0.1452	0.2432
KNN and Support Vector Machine	0.7917	0.1381	0.228

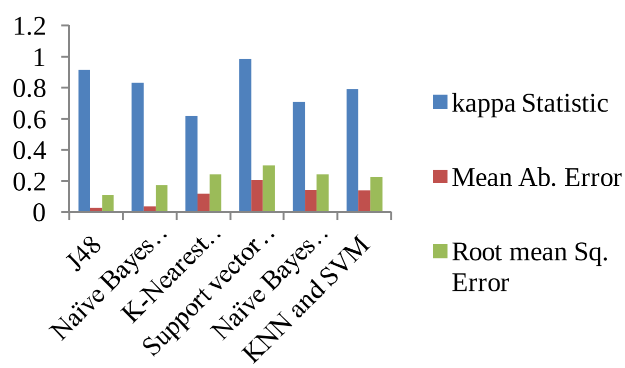

Figure 3: Bar chart shows Machine Learning Algorithms along X-axis and Kappa Statistic, MAE and RMSE values along Y-axis.

Different statistical analysis in terms of percentage of correctly and incorrectly classified instances and percentage of correctly and incorrectly classified instances for training set and testing set data for different Machine Learning Algorithms are shown in table 10 and Table 11. There bar charts are shown in figure 4 and figure 5 respectively.

The performance of our proposed work is compared or evaluated with other state of art methods for finding and detection of human emotions on social network. Tweets which are collected from twitter social network are used as the testing data set with tenfold cross validation technique. The results are reported in terms of Precision, Recall, F-score and ROC analysis summarised in table 12 [32]. The 1st advantage of our work is that no lexicon based methods are used which are restricted by their lexicons and use static prior emotion or sentiment

Table 10: Shows percentage of correctly and incorrectly classified instances of different Machine Learning Algorithms

Performance Analysis Classifier	%Tage of Correctly Classified Instances	%Tage of Incorrectly Classified Instances
J48	92.8571	7.1429
Naive Bayes Classifier	85.7143	14.2857
K-Nearest Neighbors	81.6326	18.3673
Support Vector Machine	98.5714	1.4286
Naive Bayes and J48	75.000	25.000
KNN and Support Vector Machine	82.1429	17.8571

Table 11: Shows percentage of correctly and incorrectly classified instances for training set and testing set data for different Machine Learning Algorithms

Performance Analysis Classifier	Training set		Testing set	
	%tage of correct Classified instances	%tage of incorrect Classified instances	%tage of correct Classified instances	%tage of incorrect Classified instances
J48	91.2314	8.7686	92.8571	7.1429
Naive Bayes Classifier	90.4762	9.5238	85.7143	14.2857
K-Nearest Neighbors	88.6452	11.3548	81.6326	18.3673
Support Vector Machine	98.5714	1.4286	98.5714	1.4286
Naive Bayes and J48	89.7143	10.2857	75	25
KNN and SVM	97.7143	2.2857	82.1429	17.8571

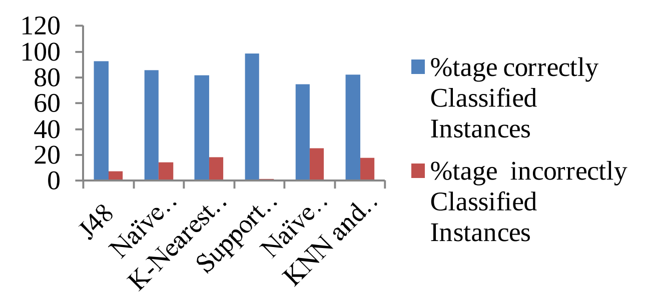

Figure 4: Bar chart shows Machine Learning Algorithms along X-axis and percentage of correctly and incorrectly classified instances along Y-axis.

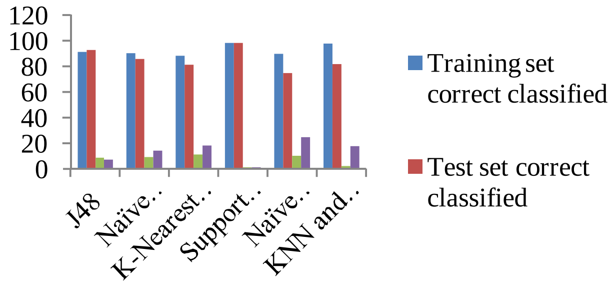

Figure 5: Correctly and incorrectly classified instances for training and test set data.

TABLE 12: Comparison of results with existing work.

S. No	Title	Method used	Precision	Recall	F-Measure	ROC Area
1	Detecting implicit expressions of emotions in text	EmotiNnet KB	48.90	42.40	45.27	NA
2	Emotion of Recognition from text based on automatically generated rules	Automatically Generated Rules	64.71	28.25	36.78	NA
3	On the identification of Emotions and author's gender in facebook	Machine Learning	32.25	41.12	33.32	NA
4	Emotion Classification using Web Blog Corpora	Machine Learning	78.30	57.33	57.33	NA
5	A Computational Approach for Analyzing and Detecting Emotions in Arabic Text	Vector Space Model (VSM)	67.50	61.68	64.30	NA
6	Sentence Level Emotion Tagging	Conditional Random Field (CRF)	67.95	65.11	60.47	NA
7	Emotions from Text: machine learning for text-based emotion prediction	Machine Learning	64.0	75.0	69.0	NA
8	Use of Porter Stemming Algo. And SVM for emotion extraction from News Headlines	Machine Learning	59.38	76.25	71.64	NA
9	**Proposed Study Results**	**Machine Learning**	**92.2**	**86.1**	**85.4**	**98.5**

values of terms regardless of their contexts. Second advantage is that the flexibility of our approach is high as compare to other corpus and lexicon based methods. The third advantage is that the machine learning algorithms are not dependent on words or syntactical features as in case of corpus and lexicon based methods.

6 CONCLUSION AND FUTURE WORK

In this paper the performance analysis of four classification algorithms in machine learning are used for the detection of human emotions on social networks and their over fitting problem analysis is being done. The four machine learning algorithms used for text classification for identifying the emotion classes are Naïve Bayes, J48, K Nearest Neighbors (KNN) and Support Vector Machine (SVM). Our experiments show different accuracy in identifying the emotion classes for the machine learning algorithms. The detail accuracy by class are shown in table 3, table 4, table 5 and table 6. The Support Vector Machine (SVM) algorithm shows the best accuracy i.e., 98.5714 % for the test set data as they can significantly reduce the need for labeled training instances. The accuracy for the other algorithms J48, Navie Bayes and K Nearest Neighbors (KNN) is 92.8571 %, 85.7143 %, and 81.6326 % respectively. For over fitting problem analysis we combined the algorithms Navie Bayes and J48 and K Nearest Neighbors (KNN) and Support Vector Machine by voting technique [33]. The results are show in figure 4 this technique is comparatively less accurate at 75.000% for algorithms Navie Bayes and J48 and 82.1429% for K Nearest Neighbors (KNN) and Support Vector Machine. This analysis shows that the algorithms are fitting i.e., they are high accurate on training data and less accurate in predicting new data (i.e. test data). The performance comparison with the existing works is shown in table 12. This work is carried on ISEAR data set as training model and tweets collected from twitter database as testing model. WEKA 3.7 interfaces is used for result analysis done under the headings cost insensitive, cost sensitive, and statistical analysis.

For future work firstly this model can be improved, and even integrated into different systems by using different ensembleing techniques. Secondly hybrid methods can be incorporated for the recognition of human emotions with real datasets.

REFERENCES

[1]. 1 R. W. Picard "**Affective Computing**", M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321
https://doi.org/10.1511/2001.28.739

[2]. Robert Plutchik "**The Nature of Emotions**", American scientist volume 89.
https://doi.org/10.1511/2001.28.739

[3]. Paul Ekman, **"Basic Emotions"**, In T. Dalgleish and M. Power (Eds.). Handbook of Cognition and Emotion. Sussex, U.K.: John Wiley & Sons, Ltd., 1999. University of California, San Francisco, CA, USA, Chapter 3.

[4]. David Watson, David Wiese, and Jatin Vaidya ,Auke Tellegen, **"The Two General Activation Systems of Affect: Structural Findings, Evolutionary Considerations, and Psychobiological Evidence"** ,Journal ol' Personality and Social Psychology 1999, Vol. 76. No. 5, 820-K38 Copyright 1999 by the American Psychological Association. Inc. 0O22-.15I4/99/S3.0O
https://doi.org/10.1037/0022-3514.76.5.820

[5]. P. N. Johnson-laird & Keith Oatley, **"The language of emotions: An analysis of a semantic field",** Cognition & Emotion, 3:2, 81-123

https://doi.org/10.1080/02699938908408075

[6]. Carlo Strapparava, Alessandro Valitutti, Oliviero Stock **"The Affective Weight of Lexicon"**, P.No 423-426.

[7]. Kaveh Bakhtiyari & Hafizah Husain **"Fuzzy Model on Human Emotions Recognition"**, Recent Advances in Electrical and Computer Engineering ISBN: 978-1-61804-156-2.

[8]. Jianhua Tao, **"Context Based Emotion Detection from Text Input"**, National Laboratory of Pattern Recognition

[9]. Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. **"Emotion Classification Using Web Blog Corpora"**, IEEE/WIC/ACM International Conference on Web Intelligence, 2007
https://doi.org/10.1109/WI.2007.51

[10]. Taner Danisman and Adil Alpkocak. 2008. **"Feeler: Emotion Classification of Text Using Vector Space Model"**, AISB 2008 Convention, Scotland.

[11]. Carlo Strapparava and Rada Mihalcea. **"Learning to identify emotions in text"**, In SAC. 2008.
https://doi.org/10.1145/1363686.1364052

[12]. Dipankar Das and Sivaji Bandyopadhyay. **"Sentence Level Emotion Tagging"**, IEEE, 2009.

[13]. Jerome R.Bellegarda. **"Emotion Analysis Using Latent Affective Folding & Embedding"**, Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 2010.

[14]. Paul N. Bennett Susan T. Dumais Eric Horvitz **"Probabilistic Combination of Text Classifiers Using Reliability Indicators: Models and Results"**, Tampere, Finland. SIGIR'02, August 1115, 2002.
https://doi.org/10.1145/564376.564413

[15]. Bincy Thomas, Vinod P, Dhanya K A. **"Multiclass Emotion Extraction from Sentences"**, IJSER0-2014.

[16]. F.M. Nazmul Haque Nahin, Jawad Mohammad Alam, Hasan Mahmuda & Kamrul Hasan, **"Identifying emotion by keystroke dynamics and text pattern analysis"**. Behaviour & Information Technology , 2014 Vol. 33,No 9, 987-996.
https://doi.org/10.1080/0144929X.2014.907343

[17]. Foteini Agrafioti, Dimitrios Hatzinakos and Adam K. Anderson **"ECG Pattern Analysis for Emotion Detection."** IEEE Transactions On Affective Computing, Vol. 3, No. 1, January-March 2012.
https://doi.org/10.1109/T-AFFC.2011.28

[18]. Caryl Charlene Escolar-Jimenez, DBA , Kichie Matsuzaki, Reggie C. Gustilo, **"A Neural-Fuzzy Network Approach to Employee Performance Evaluation"**, International Journal of Advanced Trends in Computer Science and Engineering, ISSN 2278-3091, Volume 8, No. 3, May-June 2019.

[19]. Jay Robert B. Del Rosario, "**Development of a Face Recognition System Using Deep Convolutional Neural Network in a Multi-view Vision Environment"**, International Journal of Advanced Trends in Computer Science and Engineering, ISSN 2278-3091, Volume 8, No. 3, May-June 2019.

[20]. Tina R. Patil, Mrs. S. S. Sherekar **"Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification"**, International Journal Of Computer Science And Applications, Vol. 6, No.2, Apr 2013.

[21]. Tanu Verma, Renu and Deepti Gaur, **"Tokenization and Filtering Process in RapidMiner"**, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 7– No. 2, April 2014 – www.ijais.org

[22]. Rohit Arora, Suman, **"Comparative Analysis of Classification Algorithms on Different Datasets using WEKA"**, International Journal of Computer Applications (0975 – 8887) Volume 54– No.13, September 2012

[23]. M. Ikonomakis, S. Kotsiantis and V. Tampakas **"Text Classification Using Machine Learning Techniques"**, WSEAS transactions on computers, Issue 8, pp. 966-974 Volume 4, August 2005.

[24]. Aakash Tiwari, Aditya Prakash **"Improving classification of J48 algorithm using bagging, boosting and blending ensemble methods on SONAR dataset using WEKA"**, International Journal of Engineering and Technical Research (IJETR) ISSN: 2321-0869, Volume-2, Issue-9, September 2014.

[25]. Vandana korde, **"Text Classification And Classifiers: A Survey"**, Iinternational Journal Of Artificial Intelligence & Applications (IJAIA), vol.3, No.2, March 2012.

[26]. Shweta Srivastava, **"Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining"**, International Journal of Computer Applications (0975 – 8887) Volume 88 – No.10, February 2014.

[27]. M. Trivedi, N. Soni, S. Sharma, S. Nair, **"Comparison of Text Classification Algorithms",** International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue, 02, February-2015.

[28]. E. Bhuvaneswari, V. R. Sarma Dhulipala, **"The Study and Analysis of Classification Algorithm for Animal Kingdom Dataset"**, Information Engineering Volume 2 Issue 1, March 2013.

[29]. Saima Aman, **"Recognizing Emotions in Text Thesis submitted"**.

[30]. S.B.Kotsiantis "**Supervised Machine Learning: A Review of Classification Techniques"**, Emerging Artificial Intelligence in Computer Engg., I.Magloginnies et.al. (Eds.), IOS Press 2007.

[31]. Philip K. Chan and Salvatore J. Stolfo, **"A comparative Evaluation of Voting and Meta-learning on Partitioned Data"**, Article May 1995, Source CiteSee.

[32]. S. B. Kotsiantis, I. D. Zaharakis, P. E. Pintelas, " **Machine Learning: A Review of Classification Techniques"**, Artif Intell Rev(2006)26:159-190. Published online 10 November 2007. Springer + Business Media. B.V.2007

[33]. S.K.Trisal, A.Kaul. **"K-RCC: A Novel Approach to Reduce the Computational Complexity of KNN Algorithm for Detecting Human Behavior on Social Networks"**, Journal of Intelligent and Fuzzy systems, ISSN-1064-1246, Vol. 36, No 6, pp 5475-5497, 2019.

[34]. Document, "Support Vector Machines (SVM) Regression."

[35]. www.cs.waikato.ac.nz/ml/weka/ for WEKA open source download.

[36]. https://en .wikipedia.org/