

# **Sentiment Analysis of Emoticon Based Neuro Fuzzy System**

A MAJOR PROJECT II REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF DEGREE

OF

MASTER OF TECHNOLOGY

IN

**INFORMATION SYSTEMS**

Submitted by:

**SOURABH SHRESHTH**

**2K18/ISY/12**

Under the supervision of

**Ms. Ritu Agarwal**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

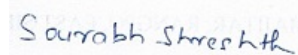
Bawana Road, Delhi-110042

JUNE, 2020

## CANDIDATE'S DECLARATION

I Sourabh Shreshth, Roll No. 2K18/ISY/12 student of M.Tech (INFORMATION SYSTEMS), hereby declare that the project report titled “**Sentiment Analysis of Emoticon Based Neuro Fuzzy System**” which is submitted by me to the Department of Information Technology, Delhi Technological University, Bawana road campus, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of and Degree, Diploma Associate ship, fellowship or other similar title or recognition.

Place :Delhi  
Date :29/07/2020



Sourabh Shreshth  
(2K18/ISY/12)

# CERTIFICATE

I hereby certify that the Major Project II report titled “**Sentiment Analysis of Emoticon Based Neuro Fuzzy System**” which is submitted by Sourabh Shreshth, Roll No. 2K18/ISY/12, Department of Information Technology, Delhi Technological University, Bawana road campus, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

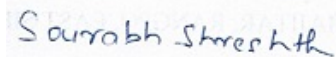
Place : Delhi  
Date : 29 July 2020



Ms Ritu Agarwal  
(Supervisor)

# ACKNOWLEDGEMENT

The success of a Major II project requires help and contribution from numerous individuals and the organization. Writing the report of this project work gives me an opportunity to express my gratitude to everyone who has helped in shaping up the outcome of the project. I express my heartfelt gratitude to my project guide Ms. Ritu Agarwal for giving me an opportunity to do my Major II project work under his guidance. Her constant support and encouragement has made me realize that it is the process of learning which weighs more than the end result. I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work. My guide helped me throughout by giving new ideas, providing necessary information and pushing me forward to complete the work.



Sourabh Shreshth  
(2K18/ISY/12)

## **Abstract**

Virtual societal space has become an integral part of human life, in which opinion generated by others plays a significant role in one's actions. Tweets as a point of the source have emerged as a way of expressing users' views on a particular subject. Historically sentiment analysis of tweets focused on polarized views, but in the real world, the opinion on a topic usually quite diverse. Therefore, using the classical classification methods cannot adequately fit the human sentiment, and we require a robust way that can match the broad scale of opinion-making.

Present work describes an approach that can fit the opinion-making scale of human sentiment by using a fuzzy neuro system to train the model for opinion-making. Emoticon used in tweets provides an additional layer with classical textual analysis to incorporate complex emotions that can be easily conveyed by emoticons. The scored sentiment is then classified using various machine learning algorithms to measure its accuracy.

Key words: Neuro fuzzy system, Sentiment analysis, Accuracy.

# Contents

S.No		Page.No
1	Chapter-1 Introduction 1.1 Motivation 1.2 Objective of the Report 1.3 Literature Review 1.4 Organization of the Report	1 2 2 2 4
2	Chapter-2 Classification algorithm for sentiment analysis 2.1 Background 2.1.1 Natural language processing 2.2 K-means 2.3 Fuzzy C-Means 2.4 Neuro Fuzzy system 2.4.1 ANFIS system	5 5 5 6 7 8 8
3	Chapter-3 Emoticon based neuro fuzzy system 3.1 Data pre-processing 3.2 Sentiment score generation 3.2.1 Scoring parameters 3.3 Neuro fuzzy based classifier	13 13 14 15 17
4	Chapter-4 Results and Discussion 4.1 Evaluation metrics 4.2 Data collection and pre-processing of data 4.3 Sentiment score generation 4.4 Classifier performance 4.4.1 K-means for classification 4.4.2 ANFIS for classification	20 20 20 22 23 24 27
5	Chapter-5 Conclusion and Future scope	31
6	References	32

## List of Figures

<b>Figure No.</b>	<b>Figure Name</b>	<b>Page No.</b>
01	Natural Language Processing	06
02	Layered ANFIS system	09
03	Pre-processing of tweets	14
04	Sentiment score generation flowchart	17
05	Traning and testing an ANFIS system	18
06	Emoticon based neuro fuzzy system	19
07	Confusion matrix	20
08	Raw data	21
09	Pre-processed final data	22
10	Sentiment score	23
11	Data distribution with cluster centroid K-means	24
12	K-means confusion matrix	25
13	K-means ROC curve	26
14	Data distribution with cluster centroid FCM	27
15	ANFIS confusion matrix	28
16	ANFIS ROC curve	29
17	Input membership function and output membership function	30

## List of Tables

<b>Table No.</b>	<b>Table Name</b>	<b>Page No.</b>
1	Emoji sentiment score	15
2	Two pass hybrid learning algorithm for ANFIS	19
3	Centroids	23
4	K-mean classification report	25
5	ANFIS classification report	28
6	Classifier accuracy	30



# **Chapter 1**

## **INTRODUCTION**

Social media has brought a colossal change in the way we build an opinion about a subject matter. With current 4.2 billion internet users around the world, approximately 3.03 billion users are active on social media[1]. Therefore building a method to model opinion-making of social media users can play a huge role in various aspects like generating a recommendation system, targeted ad delivery and many more. In this work, we are going to use tweets from twitter as our source of information for building our classification model. Twitter clocks in over an excess 600 million tweets per day with more than 300 million active users [1]. Users post their ideology in form of tweets using twitter. The tweet is a great tool to express emotions' on a particular subject.

However, the biggest problem in sentiment analysis of tweet is to determine its polarity by using various textual analysis methods. Various classical methods of classification can determine the tweets as positive, negative and neutral to an extent, but they fail profoundly in front of the vagueness of human mind opinions. Tweets containing a mix of negative and positive words can be wrongly classified as neutral sentiment when we only use textual information in tweets[2]. Hence, only using textual analysis of tweet for sentiment generation of a tweet is not sufficient to understand the real opinion of tweet's author.

Therefore before we generate a classification model using a various classification algorithm, we incorporate emoticons along with textual information of tweets for generating the data for our model. Emoticons are the facial expression of the author in a digital form representing expression like love, joy, surprise, anger, sadness, fear and more[3]. Emoticon helps in better representation of the sarcastic expression. Once we have the right set of input data, we can apply various classification algorithm with a fuzzy neuro system to generate the model and compare the result with existing algorithm.

## **1.1 Motivation**

The motivation of this thesis came from the full application of sentimental information of social media data in the field of human interaction evaluation. As universe dictates how vague a human conversation can be either in virtual means or in reality, to capture the vagueness and to model it is a very challenging task. In this thesis, our main effort is to somewhat capture this vagueness by using neuro-fuzzy system and emoticon so that the model generated can excel in the context of best available methodology.

## **1.2 Objective of the report**

The key objectives of the project work rreport can be briefly noted as follows:

- Use of emoticon as a key component of preprocessing raw tweets and calculating sentiment score.
- Implementing Adaptive Neuro Fuzzy Inference System or ANFIS algorithm for sentiment analysis of input data from social media.
- In depth analysis of ANFIS parameter, rules and membership functions.
- A detailed comparison of generated model with other classifier algorithm in terms of accuracy.

## **1.3 Literature Review**

In recent decades, it has been an uphill task for the researcher to find the sentiment of Twitter data due to its nature, abbreviations, misspells, slangs and limited size[4]. Therefore before application of any classifier on a dataset, it becomes imperative to preprocess the dataset optimally. Data preprocessing plays a key role in generation of model, and a very intuitive approach of using emoticons plays a crucial role in our project work. The prevailing steps in data preprocessing include the elimination of retweets, stop words, external links, username and hashtags[5]. In addition to this, conversion of slangs to words with an equivalent meaning along with stemming of words into their root form[6]. However, even with all these steps, previous work shows that the current methods fall when we encounter a sarcastic word that falls above or below the neutral zone[7]. Emoticon provides a great tool to combat this problem. Emoticon comprises 10% of the entire twitter traffic, but their significance is much more considerable than that[5]. Sentiment analysis by textual tools can consider sentiment as

positive, neutral and negative but incorporating emoticon stretch this category. The sentiment expressed by emoticons is reliable indicators of sentimental polarity, while there is also a significant variation in how people express emotions through emoticons and how they interpret sentiments conveyed by emoticons[8]. Extreme caution should, therefore, be exercised when using rich information in emoticons for better analysis of feelings.

Traditionally classification models were generated using classifiers like naive Bayes, decision trees, linear regression and more which are considered as hard computation based classification methods[9]. One of the most significant drawbacks of this method is the consideration of data as linear, but the data is usually non-linear. Therefore when we generate a classifier model using hard computation based classification method, the accuracy of the classifier usually is lackluster. In comparison, soft computation based method does not depend on the structure of input data and as a result, performs exceedingly well then their counterpart[10].

Fuzzy logic has shown the capability of generating continuous function to the required accuracy. On the measure of performance both neural network and fuzzy logic attains similar stature. The research in merging of various existing methodologies for the generation of classification models has been extensively carried out. In many classifier designs, we see the use of a combination of ANFIS algorithm with other approaches like genetic algorithm, Recursive Least Square algorithm and more.[11]. This allows of integration of different approaches in the training of data in the classification model. This multi-prognostic approach is well reflected by many classification methods that falls in hard computation based as the result of this approach makes classification accuracy substantially better [11].With these objectives in mind, in this project, we aim to use the self-adaptability of the neuro fuzzy based method in the approximation of arbitrary functions having virtue of neural network's better accuracy and the property of fuzzy network to model data without any specifics of the system under consideration by the use of a neuro fuzzy approach namely ANFIS system. ANFIS model based on emoticon has excellent potential to overcome the problems discussed before and generate a model that can determine the sentiment of Twitter data accurately. On the extension combination of emoticon with the textual content of tweet for generating sentiment score extend the horizon of sentiment that can be mapped in the classification model.

## **1.4 Organization of the report**

The project work report accounts the use of proposed methodology in sentiment analysis domain. The organization of project work report is given below:

1. Chapter one presents the concept explored in this work. The main aim of the chapter is to give a brief idea about the work we have done in our project.
2. Chapter two presents the background of this work.
3. Chapter three describes the subsystem of our work and their technical aspect.
4. Chapter four provides the results obtained through the code for our sentiment analysis models.
- 5 Chapter five concludes and presents suggestions for future work.

## **Chapter 2**

# **CLASSIFICATION ALGORITHM FOR SENTIMENT ANALYSIS**

## **2.1 Background**

For building a model to analyze the sentiment of tweets, we first need to make sure the data is right to generate the desired result. For that most widely known techniques use natural language processing or NLP for preprocessing of unfiltered data. Then we can illustrate the classifiers.

### **2.1.1 Natural Language Processing**

Human communicates with one another with the use of language. In the era of internet, we rely on social media for communication using website twitter, Facebook and more. Therefore for sentiment analysis of human communication in social media, we need to process the information exchange between the users of social media. Natural language processing is a technique devised to interpret the communication over the computer exchanged from one person to another[12]. Overtime after many studies, we have reached a point where we see a lot of translation application to work courtesy of natural language processing. Natural language processing is divided into phases or logical steps as input sentence, morphological processing, pragmatic analysis, semantic analysis and target representation[13].

We commence natural language processing by reading textual information from an input dataset of the sentence. Next, we break the input sentences into smaller chunks known as token containing sentence, words and paragraphs logically, and the procedure is termed as morphological processing[13]. Once we have chunks of the sentence in the next step, natural language processing establishes a relation between chunks based on syntax[14]. The current phase of natural language processing is called as syntactic analysis. Next, with the possible syntactical meaning of sentences natural language processing steps further in the analysis of the exact meaning of syntactically related chunks[15]. The step is called semantic analysis. In the next step, natural language processing pieces together all the related entities of the sentence to generate the event in a sentence in the chronologically correct way[16]. This

process is called as pragmatic analysis. All the above described working in order is termed as natural language processing and can be visually interpreted in below diagram.

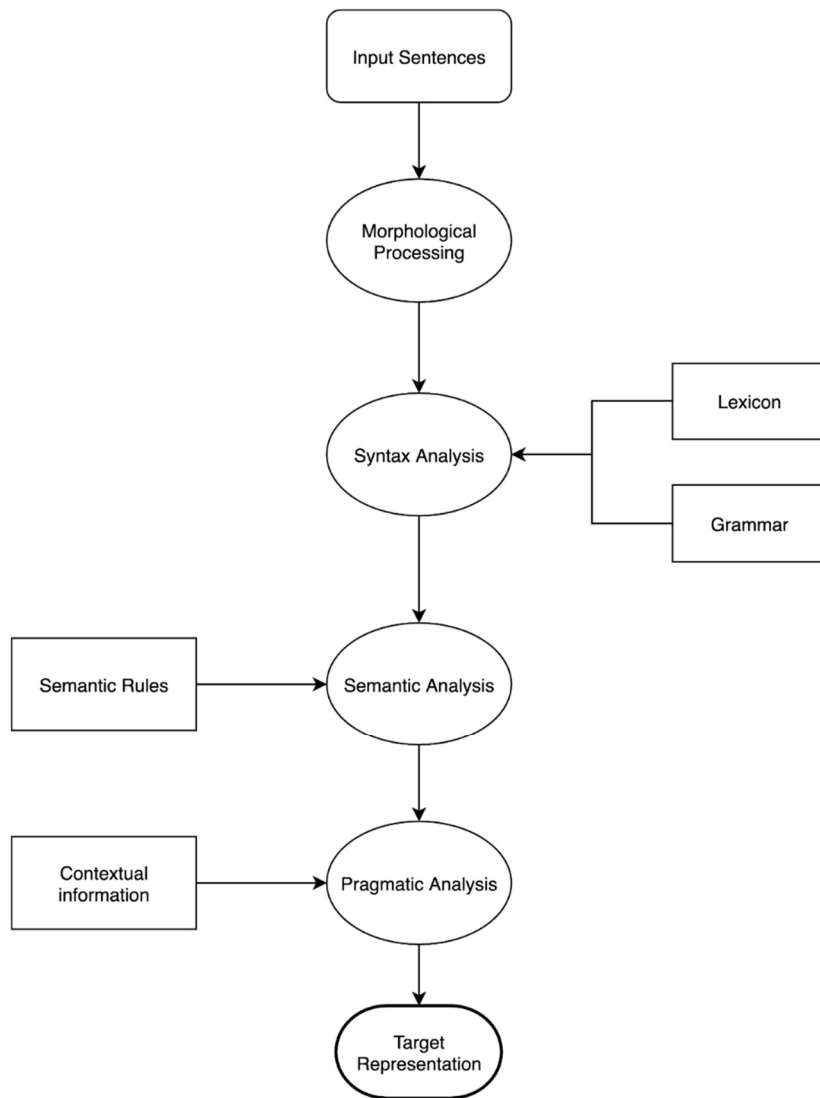


Fig 1: Natural Language processing

## 2.2 K-MEANS

K-means algorithm is originally created as a clustering algorithm that works well in domain of classification problem. K-means algorithm proceed by trying to divide an input dataset into k no of non-overlapping subgroups also called as a cluster[5]. K-mean algorithm works on the principle of minimizing intracluster differences among data points and maximizing intracluster differences. The algorithm achieves this populating cluster with data points in which the distance between cluster centroid and data points is minimum.

The K-mean algorithm works as follows:

- (1) Cluster centroid are chosen at random initially.
- (2) For each data point calculate its distance with every cluster centroid.
- (3) On the basis of proximity between data points and cluster center assign data point to a single cluster center.
- (4) Centroid of a cluster is updated by equation(2.1)

$$V_i = \sum_{j=1}^{n_i} \frac{X_{ij}}{n_i} \quad (2.1)$$

Where V is the cluster center, X is dataset,  $1 \leq i \leq c$ ; c denotes total number of cluster.

- (5) Updated cluster center. Now using the new cluster center recalculate the distance of data points and all cluster center.
- (6) If no reassignment of data point to cluster center is reported then stop else repeat step 3 to 5 iteratively.

### 2.3 Fuzzy C-Means

Fuzzy C-Means or FCM is an unsupervised clustering algorithm that is widely used for applications like classifier design, clustering and more. The algorithm calculates the distance between input data points wherein this project we have used emoticons and form cluster around central points[17]. As a result data point within a cluster have higher similarity with each other than the data points in different cluster.

Fuzzy C-Means algorithm works in following way:

- (1) Initialize  $P = [p_{ij}]$  matrix,  $P(0)$
- (2) For every succeeding k-step: calculate the centroid vectors  $C^k = [c_j] P^k$

Where j is the centroid iteration k denotes data iteration number, c denotes number of clusters.

(3) Update  $P^k, P^{k+1}$

$$p_{ij} = \frac{1}{\sum_{k=1}^C \frac{x_i - c_j}{x_i - c_k}} * \frac{1}{m-1} \quad (2.2)$$

Where m is the membership function.

(4) If  $P^{k+1} - P^k < \epsilon$  then STOP; otherwise return to step2.

Here  $\epsilon$  is the termination criterion i.e. the minimum difference between the cluster centroid.

## 2.4 Neuro fuzzy System

A Neuro-Fuzzy system is defined as a network of hybrid structure having two components neural network and fuzzy logic with a unique structure and lack of any feedback mechanism comprising of weights, non-fuzzy signal and activation functions[10]. The main idea behind the modelling of such a system is to determine the parameters for input and output membership function that can fit a fuzzy logic based inference system. For this purpose, the neural network procedures are used to obtain the desired parameters. Of all available fuzzy rule based system three among them are most popular namely Mamdani, Sugeno, Tsukamoto . In this project we are using ANFIS system having a five layered neural network structure.

### 2.4.1 Adaptive Neuro Fuzzy Inference System

In this project work, Adaptive Network-based Fuzzy Inference System or ANFIS in short is used as a key algorithm in generation of classification model. Adaptive Fuzzy Inference System or ANFIS is a neuro fuzzy based classification technique which is by looking at terminology itself echoes an algorithm which amalgam of fuzzy system and neural network [18]. In ANFIS, both parameters antecedent and consequent is adapted in such a way that the ANFIS architecture combines the working of both Sugeno and Tsukamoto models. ANFIS approach is a combination of a fuzzy system and a layered neural network where primary role of the fuzzy network is to handle the uncertainty of system architecture in consideration to reduce error. Also, the neural network provides s the adaption strength to the parameters in use. Using this hybrid method, we generate a fuzzy model with its input variables initially, which is derived from the rules extracted from the training data of the system. Next, the



neural network fine-tunes the initial fuzzy model rule base in order to generate the system's final ANFIS configuration. ANFIS is used as the basis in this project define real-world structures.

### 2.4.1.1 ANFIS Structure

A fuzzy inference system of Sugeno's and Takagi type with a single output rule and two inputs looks like:

if  $x$  denotes  $A$  and  $y$  denotes  $B$  then  $z$  denotes  $f(x,y)$

Where  $z=f(x, y)$  denotes a crisp function, and  $A$  and  $B$  are ANFIS system fuzzy sets.. When  $f(x, y)$  is constant, it leads to the formation of Sugeno fuzzy model of zero order where a fuzzy singleton specifies each rule of consequence[19]. A fuzzy inference method proposed by Sugeno also termed as type 3 is shown in figure below. Here we can see that at every node its output is the resultant of input variables boosted by a constant [20].The resultant of the system is the weighted average of every rule from rule base output. A diagram of layered structure of ANFIS system is shown below.

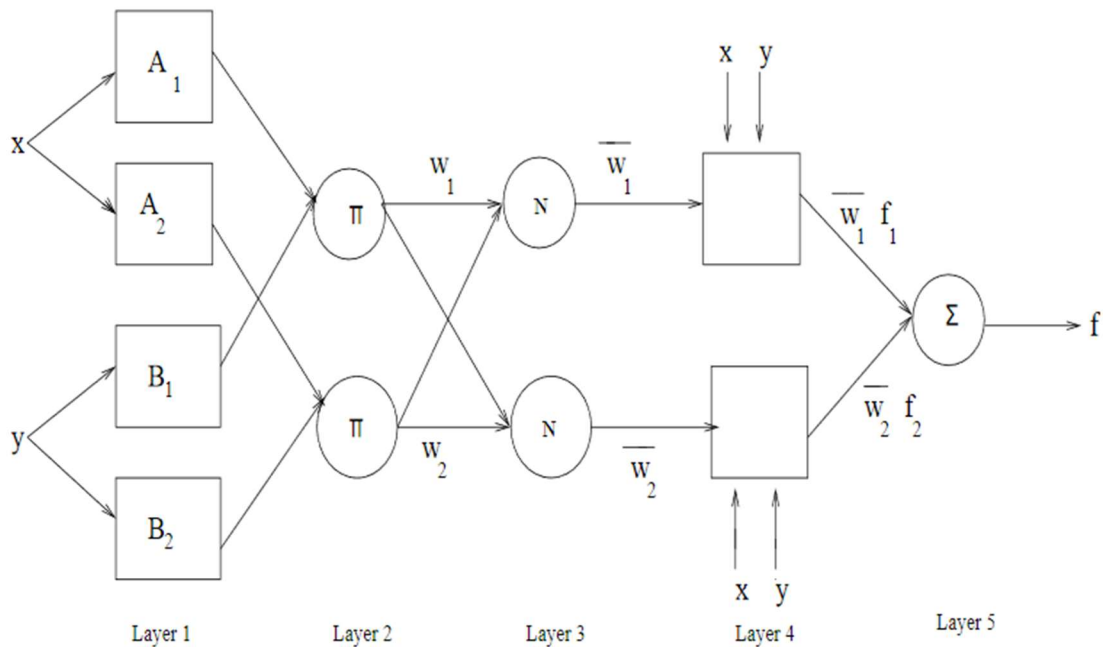


Fig 2: Layered ANFIS system

Each layer in a general ANFIS structure is described as follow:

Layer 1: All nodes in layer one of an ANFIS structure is adapted with function:

$$Output_a^1 = \mu_{Anfis_i}(z) \quad (2.3)$$

Where,  $z$  is the input for a node,  $Anfis_a$  is the variable termed as linguistic variable in association with the input node function and  $\mu_{Anfis_i}$  denotes a function called as the membership function of  $Anfis_a$ .  $\mu_{Anfis_a}(z)$  is expressed as:

$$\mu_{Anfis_a}(z) = \frac{1}{1 + \left[ \left( \frac{z - o_i}{m_i} \right) \times \left( \frac{z - o_i}{m_i} \right) \right]^{n_i}} \quad (2.4)$$

Where  $z$  represents input and  $\{m_i, n_i, o_i\}$  is the parameter set of premise type.

Layer 2: The layer two of ANFIS represents a node with fixed value which estimates the rule's firing strength denoted by  $f_a$ . The layer 2 of ANFIS is represented as:

$$Output_a^2 = f_a = \mu_{Anfis_a}(x) \times \mu_{\beta_a}(y) \quad (2.5)$$

Where  $a = 1, 2$ .

Layer 3: Each node in layer three is a set of node of fixed type. Each  $a^{th}$  node in layer three represents the ratio of fire strength of their rule to the combined aggregate of fire strengths of every rule of the system. The output of layer for the  $a^{th}$  node is given by:

$$Output_a^3 = \bar{f}_a = \frac{f_a}{f_1 + f_2} \quad (2.6)$$

Where  $a = 1, 2$ .

Layer4: Every node in layer four is node of adaptive type with function represented as:

$$Output_a^4 = \bar{f}_a g_a = \bar{f}_a (l_a x + m_a y + n_a), a = 1, 2$$

Here  $\bar{f}_a$  is layer 3 output and  $\{l_a, m_a, n_a\}$  is the corresponding set of parameters set. (2.7)

Layer 5: Fifth layer of ANFIS structure represents the overall output of the system.

$$Output_a^5 = \text{Final output} = \sum \bar{f}_a g_a \frac{\sum_i f_a g_a}{\sum_i f_a} \quad (2.8)$$

### Learning Algorithm:

ANFIS structure's final result can be formulated as a linear combination of the corresponding parameters. The output  $o$  can be formulated as:

$$\begin{aligned} o &= \frac{f_1}{f_1+f_2} o_1 + \frac{f_2}{f_1+f_2} o_2 \\ &= \bar{f}_1 o_1 + \bar{f}_2 o_2 \\ &= (\bar{f}_1 a) l_1 + (\bar{f}_1 b) m_1 + (\bar{f}_1) n_1 + (\bar{f}_2 a) l_2 + \\ &\quad (\bar{f}_2 b) m_2 + (\bar{f}_2) n_2 \end{aligned} \quad (2.9)$$

Where consequent parameters are  $(l_1, m_1, n_1, l_2, m_2, n_2)$ .

The antecedent parameter and the consequent parameter in an ANFIS structure is selected by the user. A combination of the parameters algorithm is shown in table 2.

Raw data input preprocessing is an essential step to alleviate the accuracy of classification. The traditional method like Word2vec, TF-IDF and more was sufficient for sentiment classification till the polarity of classification is limited to positive, negative and neutral[21]. However, with the requirement of a broader classification methodology and increase in accuracy of classification, we require a method that can minimize the preprocessing error. One of the recent work in this approach is the conversion of textual information in numerical data by taking emoticon as a base of conversion[21]. Various type of emoticons is in use like

an image-based emoticon, text-based emoticon and more, out of which image-based emoticon or emoji is more prevalent than others[22]. However, even changing the preprocessing method, model' accuracy is only increased by meagre 1-4 %. Therefore the next step of improvement can come from by making a change in classifier itself. Recent work in this way has shown a great promise by using a neuro-fuzzy classifier. Out of the comparison between the various neuro-fuzzy methods, a combination of ANFIS classifier and FCM clustering has shown significant improvement in increasing the accuracy of the classifier by 3-7% by using standard preprocessing methods only[23]. Therefore a combination of neuro-fuzzy method and emoticon-based preprocessing is suitable to increase the accuracy of classification. In our work, we are exploring the idea of a hybrid method for sentiment analysis of twitter data and comparing the result with a traditional classifier.

## **Chapter 3**

### **EMOTICON BASED NEURO FUZZY SYSTEM**

The project work primarily shows the use of a neuro-fuzzy system in the sentimental analysis of twitter data using emoticon as a core ingredient of the data set and compare the performance with other classifiers. The first step is the data collection of twitter data using tweepy. Tweepy allows us to extract tweets based on our requirements. Here in this project as we are using emoticons, we can use query to search for tweets using emoticons values. For example, \U0001F602 refers to happy, 1F62D = sad, 1F621= angry, 2764 = love, 1F61C = playful, 1F631 = confused and more. The methodology used can be primarily divided into three steps data preprocessing, Sentiment score generation and neuro fuzzy classifier based model generation.

#### **3.1 Data Preprocessing**

The tweets collected using tweepy from twitter website contains noisy data that affects the accuracy of classifier adversely. Noisy data refers to the data which contains in necessary information that affects system performance. When we collect data using tweepy various noises are username, retweets, slangs and more. Therefore we can apply following techniques for cleaning of raw tweets.

In the preprocessing of data we first have to check null value and missing value in our dataset to make sure our dataset is consistent. We will replace them with a blank space since it might be a \_ or - or punctuation with no space from the next word, and we don't want the words to join together. Next step involves removing of Re-Tweets, HTTP Links,spaces and punctuations which are not relevant in evaluating the sentiment score.

Further the text available in tweets are split into sentences using Sentence Tokenizer.For maintaining accurate polarity score we have remove all sentences with less than five words and all the character in sentences are converted into lower case. For simplicity we have shown limited the no of emoticon score in sentences to six only which will be further detailed. Next we remove stop word in our input data as the sop word represents word that are frequently

used like he, they and more that do not carry sentimental analysis value. Stop word does not help in discriminating relevant and irrelevant sentences therefore we remove stop word in preprocessing of data. Finally, the remaining word are converted to their base form by a process called as stemming. Stemmed verb are helpful in understanding any tweet subjectivity. The preprocessing steps for better understanding can be illustrated in block diagram as shown below.

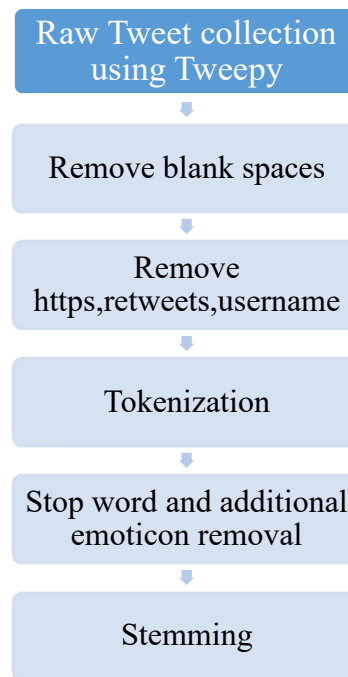


Fig 3: Preprocessing of tweets

### 3.2 Sentiment Score Generation

After preprocessing of raw tweets, we can move toward generating sentiment score. Studies conducted on emoticons have estimated intensity of emoticon in terms of integer polarity. In this project, we are selecting the six most commonly used emoticon out of 751 available emoticons for showing sentiment score rest can be visualized in the emoji ranking website[24]. The emoticon selected is represented in the table below with their score.







Emoji	Score
	0.221
	0.445
	0.746
	-0.093
	-0.173
	-0.397

Table 1: Emoji Sentiment Score

### 3.2.1 Scoring Parameters

Using emoticon as a critical feature of our dataset, we have to make sure to cover all feature of tweets. Which include general text preprocessing, hashtag and emoticons, besides we also focus on the pattern in parts of speech (POS). The advantage of adding POS to our analysis is incorporating the subjectivity of tweets in our sentiment analysis. We achieve our goal by adding stemmed verbs and noun-adjective pairs in our sentiment score generation algorithm. Till this point, we are left with emoticons, keywords and hashtag in our cleaned dataset our algorithm decide data to be shifted for each input encountered. Since we are using emoticon

score instead of integer polarity based on Emoji Sentiment Ranking an average of all emoji score is set to 0.124833 denoted as *averageChangeInSentiment*[24].

Algorithm steps for generating sentiment score are as follows:

- (1) Initialize P as sentiment score, for each emoticon encountered assign sentiment score based upon the score of emoticon in emoji ranking website.
- (2) For every hashtag, if positive increment by  $2 * averageChangeInSentiment$  else decrement by  $2 * averageChangeInSentiment$  for negative hashtag.
- (3) For POS tags, if an adjective precedes a noun then we decrement or increment by  $2 * averageChangeInSentiment$  depending upon the polarity of word else we decrement or increment by *averageChangeInSentiment*.
- (4) For POS tags, if verb is encountered we check positive or negative on the stem of word and we decrement or increment by  $2 * averageChangeInSentiment$ . For adjective repeat the above process without stemming.
- (5) Save sentiment score in text format for further process.

At the end of above described steps, the resultant textual data is ready to form as input to classifiers for generation of model that can determine the sentiment of tweets. A flowchart is described below for illustration of sentiment score generation where p represents sentiment score.



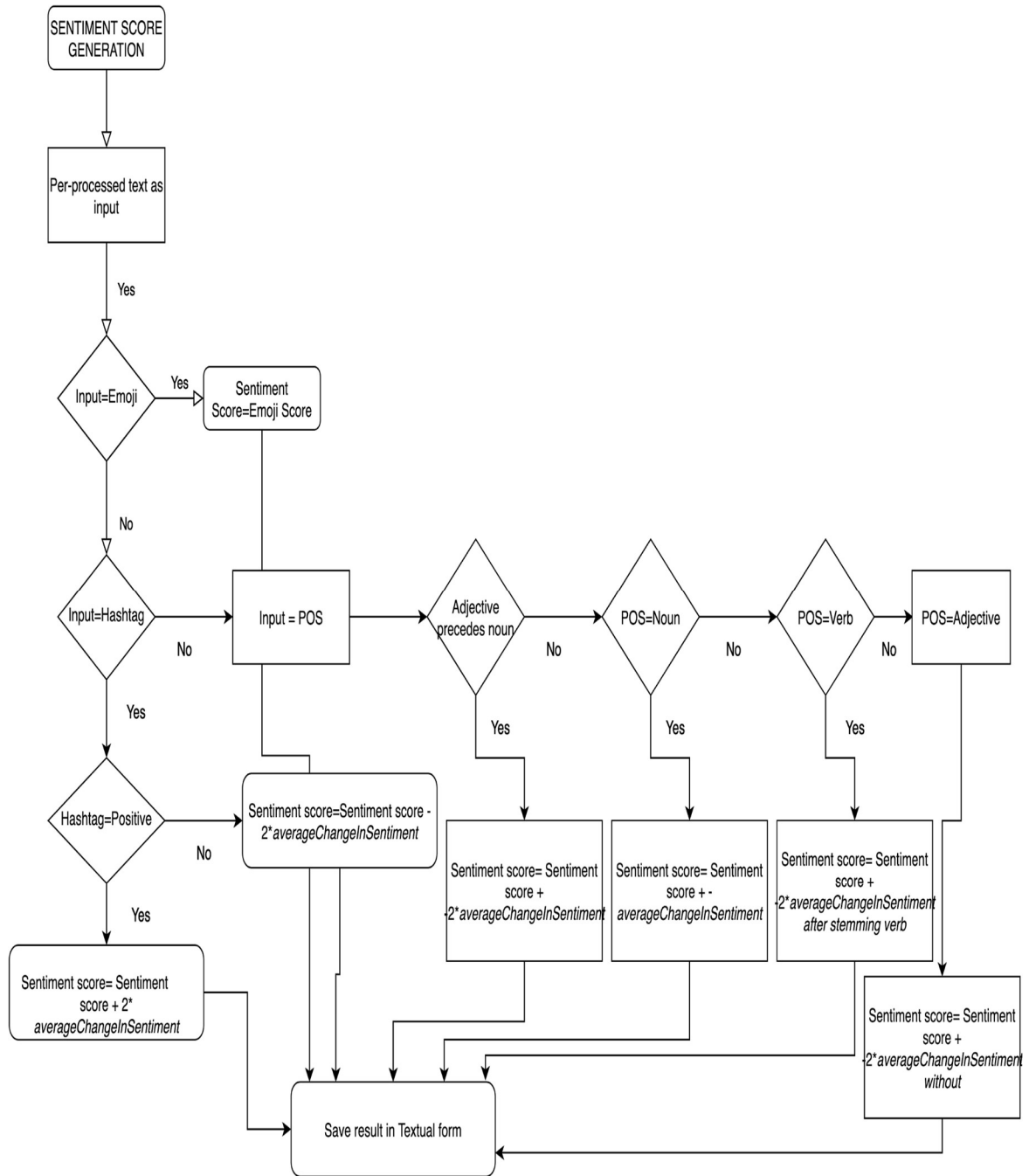


Fig 4: Sentiment Score Generation Flowchart

### 3.3 Neuro Fuzzy based Classifier

Neuro-fuzzy system integrates the advantage of learning in ANN and use of fuzzy system if-then rules membership function to a higher degree of accuracy. In this work, we are using ANFIS system to classify tweets because it has higher accuracy of forecasting and faster

convergence in comparison to other classifiers. Once we have input data ready, we can move forward in the application of ANFIS in the next step of our work. Before applying the ANFIS algorithm, we have to develop FIS, i.e. fuzzy inference system then tune it using anfis function. There is numerous FIS model like Grid Partitioning, Subtractive Clustering and more, in this project, we are using FCM. In FCM, we set a constant number of the cluster at the beginning with weights assigned randomly to them[25]. Centroids are updated by running the algorithm described previously alternatively. A flowchart of ANFIS training system is shown below.

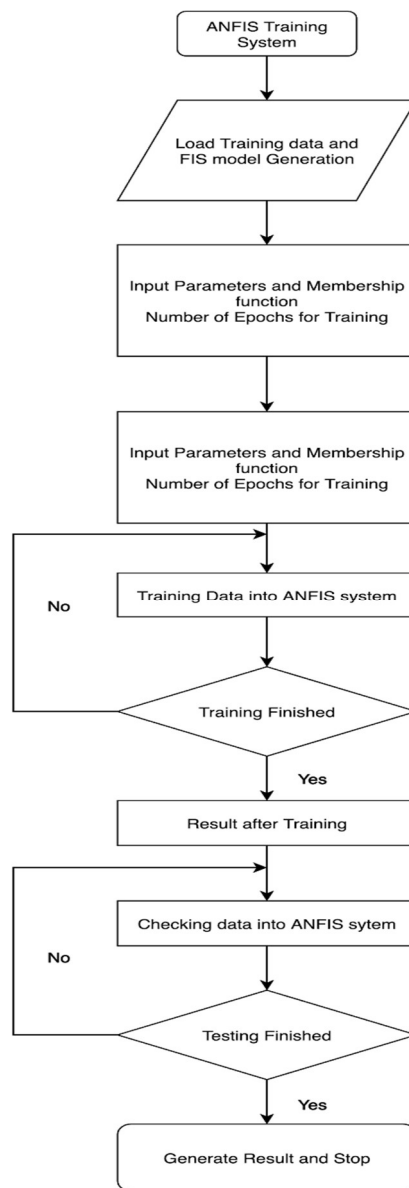


Fig 5: Training and testing an ANFIS system.

The ANFIS system training and testing methodology is shown in the above figure. The process starts with taking input and output vector generated by the FIS model and checking the pair for error. The data is split into a ratio of training and testing data which is essential for our model generation and validation. The training data generate the parameters for the membership function to minimize error generated between the resultant and desired output. Consequent parameters is calculated using least squared method. If this error generated is higher than a certain limit, then premise parameters will be modified by applying the gradient decent process. The process stops when we achieve a desirable error value.

ANFIS training learning rule in this project, we are using a hybrid learning method made of least squares method and gradient descent. A tabular representation of method is shown below:

	Forward Pass	Backward Pass
Antecedent	Fixed	Gradient Descent
Consequent	Least Square Estimation	Fixed
Signals	Node Output	Error Signal

Table 2: Two Pass Hybrid Learning Algorithm

ANFIS training and learning generates the model and model is tested with testing data. Testing data contains data not included in the training process, which is used to evaluate system performance. System performance is determined by using metrics accuracy, error curve, generated rules and output.

The emoticon based fuzzy system comprising of three sub parts namely preprocessing, sentiment score generation and neuro fuzzy based classifier is used in this project to generate a classification model that perform sentiment analysis of twitter data, A compact view of the system is shown below.

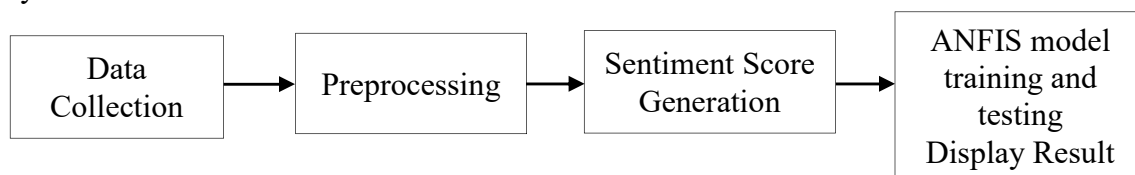


Fig 6: Emoticon Based Neuro Fuzzy System

## Chapter 4

### RESULTS AND DISCUSSION

#### 4.1 Evaluation Metrics

Evaluation metrics are known as measurement of results and outcome that consolidates the effectiveness of proposed model. The evaluation metric used in our project work are accuracy, precision, recall and f-measure. The mathematical equations are denoted in equation () where TruePositive or TP means prediction is positive, FalsePositive or FP means prediction is positive, TrueNegative or TN means prediction is false, FalseNegative or FN means prediction is negative .

$$\text{Accuracy} = \frac{\text{TruePositives} + \text{TrueNegatives}}{\text{Total Sample}}$$

$$\text{F-Measure} = \frac{2 * \text{TruePositives}}{2 * \text{TruePositives} + \text{FalsePositives} + \text{FalseNegatives}}$$

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}}$$

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNeagtives}}$$

Confusion matrix also termed as error matrix is a metric that allow visualization of model's performance. Confusion matrix is used in our results for visualization of system performance.

A confusion matrix can be interpreted as in diagram below:

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig 7: Confusion Matrix

#### 4.2 Data Collection and Preprocessing of data

Using a python library tweepy, we collected recent tweets from tweet stream and filtered the tweets based on emoticons, language and retweets. Twitter exposes a STREAM API which

allows the researcher to get access to the tweet stream. Any user can log into the developer account from their twitter account and request access to twitter. On approval, a set of API keys is made available for access. A snippet of raw data saved in csv file is shown in figure below:

1	All of these @DaleJr tweets have me like 🤔- #ForeverAFan <a href="https://t.co/jWj6HYT6gM">https://t.co/jWj6HYT6gM</a>
2	â€” 100 🤔;; Thank you for keeping my DM alive despite of me being an awkward species 🤔, lets talk more in the futureâ€” <a href="https://t.co/9uxlBNzvvh">https://t.co/9uxlBNzvvh</a>
3	@MissJL
4	@fr8ordie @TuckerCarlson @FoxNews Sorry for bringing education into it 🤔œ
5	All that pillow talking shit dead 🤔- Fr let me know some
6	All this guy does is make me laugh 🤔, @RB_Official #ninjalevels <a href="https://t.co/Q4mNviuxAb">https://t.co/Q4mNviuxAb</a>
7	All I saw was Helga rescuing Geraldâ€™s shoe 🤔, The poor boy. And Helga saying she has seen better shrines for Arnold lmao bye 🤔:🤔»
8	@_blessicaaax3 Im convinced u dont get enough 🤔• but u talkin bout me doe 🤔-
9	@imacelebrity Lovely to see Ant and Dec back again both looking as great as ever 🤔€ðŸ— looking forward to watching the celebs squirm 🤔œ
10	@DaveTodayFM Must be spooning Velcro girl so 🤔œ
11	Just wanna go back to normal 🤔•
12	Also me and my Pikachu are adorable 🤔- <a href="https://t.co/ABbz06mEgC">https://t.co/ABbz06mEgC</a>
13	Hate
14	Big Hero 6 was a GREAT movie! Omg I cried so much 🤔-
15	@quayaustralia @briannamati give my wife free glasses 🤔! please 🤔-
16	@xxmelcixx0612 @Queenhildaglam Watch this video, it will make you smile!!!! 🤔- 🤔- â€” <a href="https://t.co/82fZK1CLtd">https://t.co/82fZK1CLtd</a>
17	And i need my nails done 🤔-
18	Damn @Kenny_Rhymes aka Randy Moss ainâ€™t here this holiday 🤔•
19	@CodeineCookies I feel a way about being compared to Ricky Smiley lol 🤔!j
20	@104_glen deffo Timmy 🤔, <a href="https://t.co/GuywUhjfAt">https://t.co/GuywUhjfAt</a>
21	@AlyssaMastro44 My hair is getting so gray 🤔-
22	@acashmoney7 Looks like im not changing my number 🤔,
23	all my female cousins ghetto asl & my male cousins hood asf 🤔- & i fit right in LMAOOOO
24	#AMAs I want to watch 🤔-
25	A lot of Cowboys slander on TL 🤔•
26	@jennjames22 Just heard they wonâ€™t be coming to our house until tomorrow. So no internet or TV. 🤔•
27	I will never get the saying ignorance is bliss NEVER 🤔•
28	Also my bed and my food 🤔œ <a href="https://t.co/YileFCrrGP">https://t.co/YileFCrrGP</a>
29	@101greatgoals Heâ€™s also not saved a penalty yet but your point is 🤔,
30	I hope I don't work Thanksgiving 🤔•

Fig 8: Raw data

A total of 14123 tweets were collected and processed in preprocessing step to generate our input data set.

From input data set all tweets having less than five words, retweets, usernames and stop words are discarded in preprocessing steps. After the stemming process, we were left with nearly 5000 tweets in our final dataset. Analysis of our dataset leads to the conclusion that six emoticons shown in fig are available more than the threshold of 1200 tweets. Therefore all tweets without these six emoji are discarded, and we are left with nearly 3000 tweets. At this point, we are ready with our data and can proceed toward sentiment score generation. A snippet of preprocessed data is shown in figure below.

```

[u'hate', u'james', u'works', u'friends', u'cam', u'amp', u'just', u'nothing', u'sit', u'around']@,
[u'billing', u'coding', u'annoying', u'class', u'classes', u've', u'takeeee', u'big', u'ass', u'book', u'just', u'looking', u'codes', u'injures']@, @, @,
[u'wish', u'molly', u'comb', u'kensley', u'hair', u'least', u'show', u'camera']@,
[u'please']@,
[u'think', u'one', u'lives', u'wit']@, @, @, @,
[u'feel', u'girl']@,
[u'kinda', u'feel', u'like', u'college', u'football', u'season', u'ended', u'saturday', u'grier', u'broke', u'finger']@,
[u'gorgeous', u'picture', u'love', u'thanks', u'picture']@,
[u'bouta', u'hit', u'one', u'ex', u'like', u'yerrrrrr', u'come', u'make', u'band', u'real', u'quick', u'split', u'dc', u'f']@,
[u'contestant', u'red', u'close', u'winning']@,
[u'city', u'melbourne', u'sleep', u'well']@,
[u'fun', u'games', u'us', u'like', u'laughed', u'fact', u'someone', u'sat', u'day', u'crank', u'world', u'cup', u'defo']@,
[u'doesn', u't', u'matter', u'good', u'think', u'guy', u'bcos', u'deep', u'still', u'fucking', u'bastor']@,
[u'hopefully', u'turned', u'tho']@,
[u'm', u'proud', u'love', u'love', u'much']@, @,
[u'captions', u'killing']@,
[u'hands', u'sweating']@,
[u'two', u'kill']@,
[u'looks', u'like', u'washing', u'machine', u'finna', u'go', u'night']@, @,
[u'angry', u'litol', u'bean']@,
[u'left', u'nike', u'hat', u'someone', u'house', u'don', u't', u'know', u'want', u'back']@, @,
[u're', u'phenomenally', u'outstanding', u'just', u'way']@, |

```

Fig 9: Preprocessed final data

### 4.3 Sentiment Score Generation

Final data displayed in fig is used as input to the sentiment score generation algorithm described in fig. As soft clustering technique does not require extensive training, therefore we are only considered of incorporating all features of tweets in our sentiment score generation methodology. We used a combination of regular expression and control statement in our code to identify POS tags like noun, verb, adjective and their precedence to generate sentiment score required as an input to neuro fuzzy system. Sentiment score generated for each tweet is split into training, and testing data set on labels confused, angry, sad, happy, playful and love about which tweets are classified. Sentiment score is used to train a classification model by using k-means and ANFIS algorithm and later evaluated for efficiency. A screenshot of the sentiment score generated for our data is shown below in fig 10.

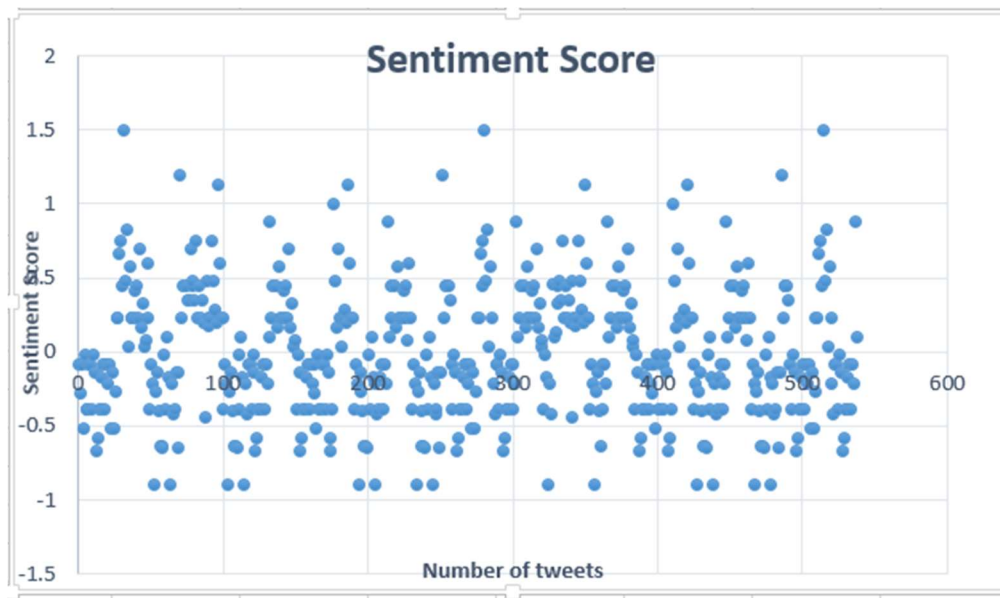


Fig 10: Sentiment score

#### 4.4 Classifier Performance

Sentiment score generated for each tweet is indexed as input in our final dataset. We run our code for all cases to get following centroid in the table below.

Emoticon	Our Algorithm	K-means
happy	0.21	0.41
playful	0.46	0.76
love	0.86	2.01
sad	-0.11	0.19
angry	-0.36	-0.13
confused	-0.64	-0.47

Table 3: Centroids

In this project, we have six labels namely happy, playful, love, sad, confused and angry to which we are classifying our data based on the proximity of sentiment score generate for

each input tweet. The model generated is evaluated against this label by using metrics confusion matrix, ROC curve and accuracy for both classifiers.

#### 4.4.1 K-means clustering for classification

For better visualization of data points a data distribution plot with black dots representing centroid for emotions namely confused, angry, sad, happy, playful and love in order in fig 11 below.

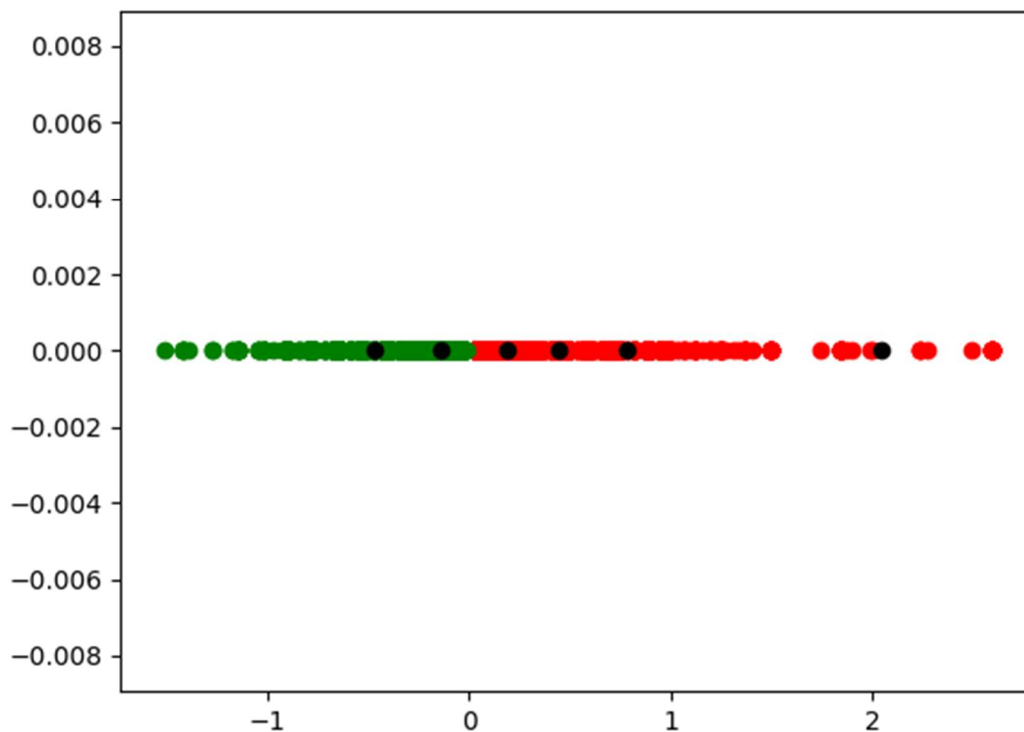


Fig 11: Data distribution with cluster centroids K-means

##### 4.4.1.1 K-means performance

K-means algorithm is evaluated using evaluation matrices in next section.



### Classification report

K-means classification report generates confusion matrix and classification report is shown in table below:

	precision	recall	F1-score	support
0	0.83	0.81	0.82	537
1	0.67	0.66	0.67	491
2	0.76	0.76	0.76	532
3	0.71	0.81	0.76	496
4	0.75	0.65	0.70	420
5	0.68	0.66	0.67	471

micro average	0.73	0.73	0.73	2947
macro average	0.73	0.73	0.73	2947
weighted average	0.74	0.73	0.73	2947

Table 4: K-means classification report

### Confusion Matrix

K-means confusion matrix based on classification report is shown in figure below:

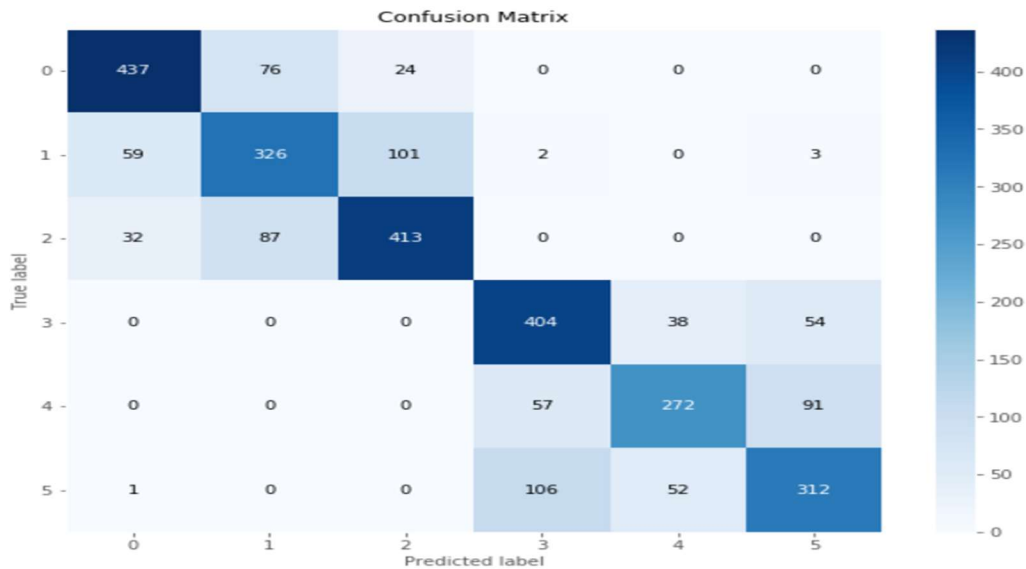


Fig 12: K-means confusion matrix

## ROC Curve

K-means ROC curves for analysis of model performance on dataset is shown in figure below:

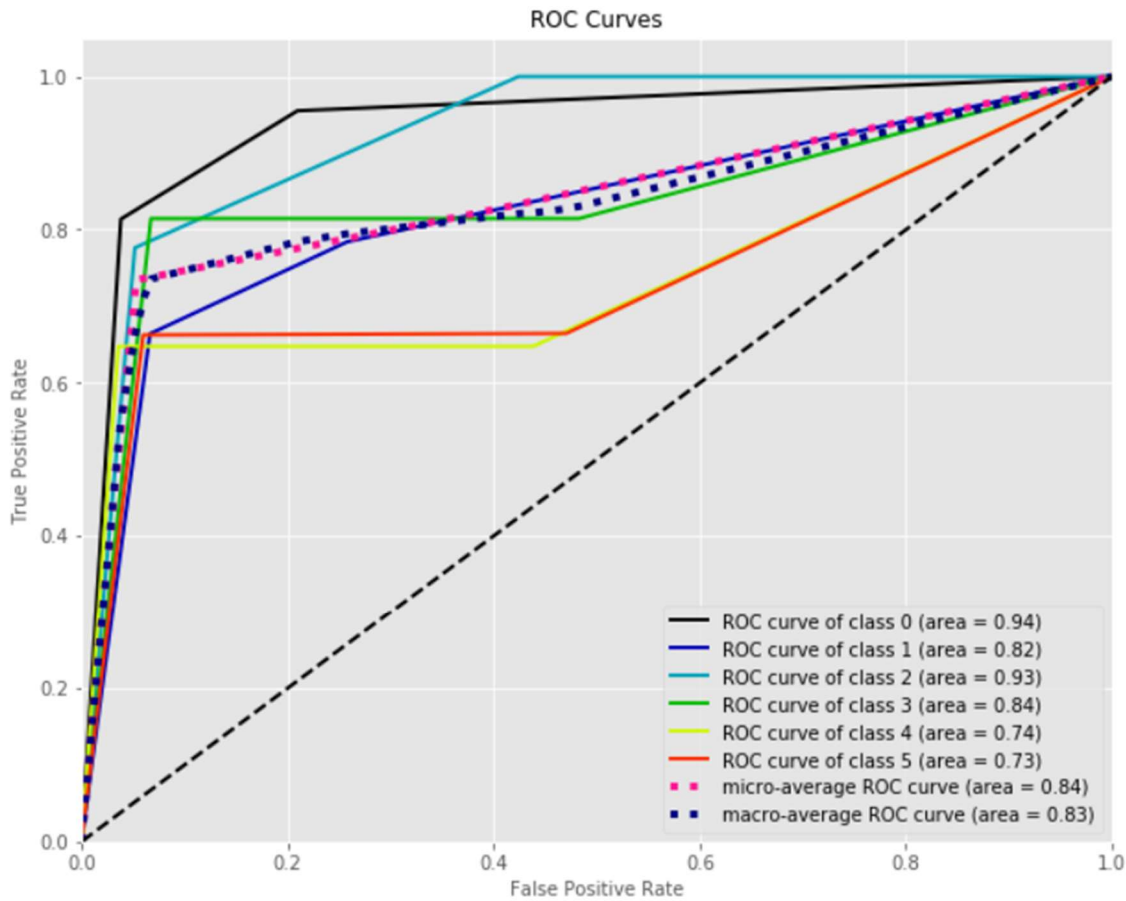


Fig 13: K-means ROC curve

### 4.4.1.2 K-means algorithm Analysis

K-means requires non-overlapping of clusters which as expected leads to a fall in the performance of the classifier. From the data distribution with a cluster centroid figure and centroid center table, we can see that K-means is very close to the desired prediction of confused and angry sentiment but perform horrendously for other sentiments. The evaluated overall accuracy of K-means classifier is 73.41% which is near to the estimated value for our dataset. The main reason for the lousy accuracy of k-means boils down to clustering technique, unstructured dataset, lack of incorporating subjectivity of tweets instead of focusing tweet's objectivity only. Some of these shortcoming are handled by neuro fuzzy classifier in our project.

#### 4.4.2 ANFIS for classification

In the ANFIS training system shown in fig, we used FCM clustering for fis generation. For better visualization of data points a data distribution plot with black dots representing centroid for emotions, namely confused, angry, sad, happy, playful and love in order in fig 14 below using fcm clustering.

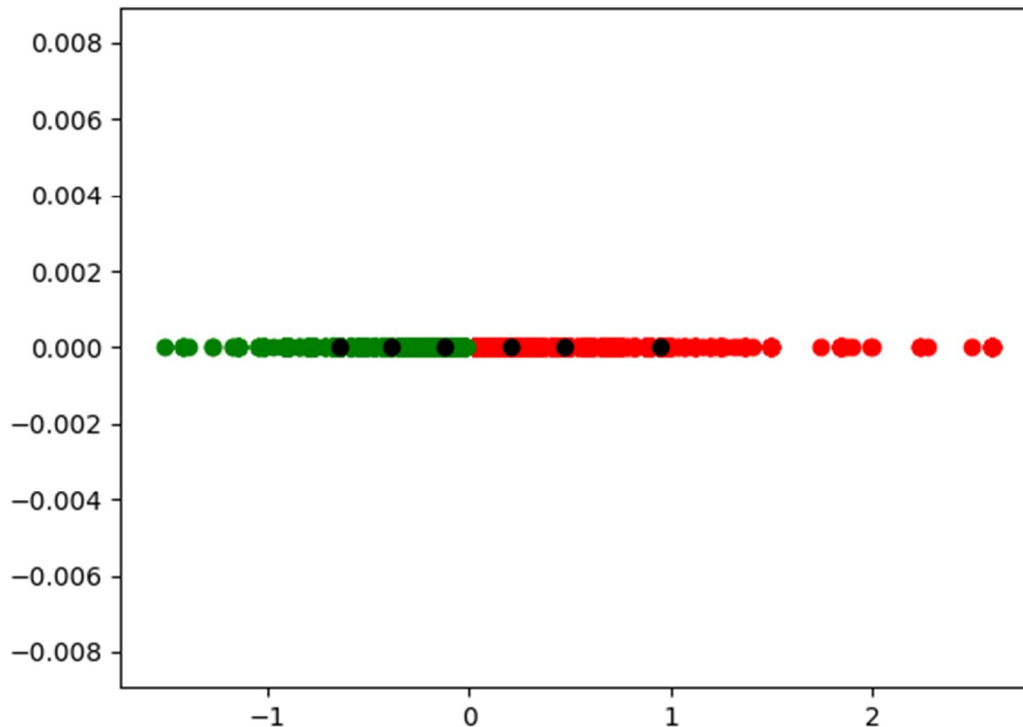


Fig 14: Data distribution with cluster centroids FCM

##### 4.4.2.1 Evaluation of ANFIS performance

We used genfis3 or fuzzy c-means for FIS creation in ANFIS training system. FIS object generated is then exported to the ANFIS system for tuning of our classifier. RMSE is the de facto metric for measure ANFIS performance, but for the sake of simplicity, we used accuracy metric used in other classifiers for evaluation of classifier

### Classification report

ANFIS classification report generates confusion matrix. Classification report is shown in table below.

	precision	recall	F1-score	support
0	1.00	1.00	1.00	537
1	0.88	0.80	0.84	491
2	0.83	0.90	0.86	532
3	0.86	0.96	0.91	496
4	0.93	0.87	0.90	420
5	0.89	0.83	0.86	471
micro average	0.90	0.90	0.90	2947
macro average	0.90	0.89	0.90	2947
weighted average	0.90	0.90	0.90	2947

Table 5: ANFIS classification report

### Confusion Matrix

ANFIS confusion matrix based on classification report is shown in figure below:

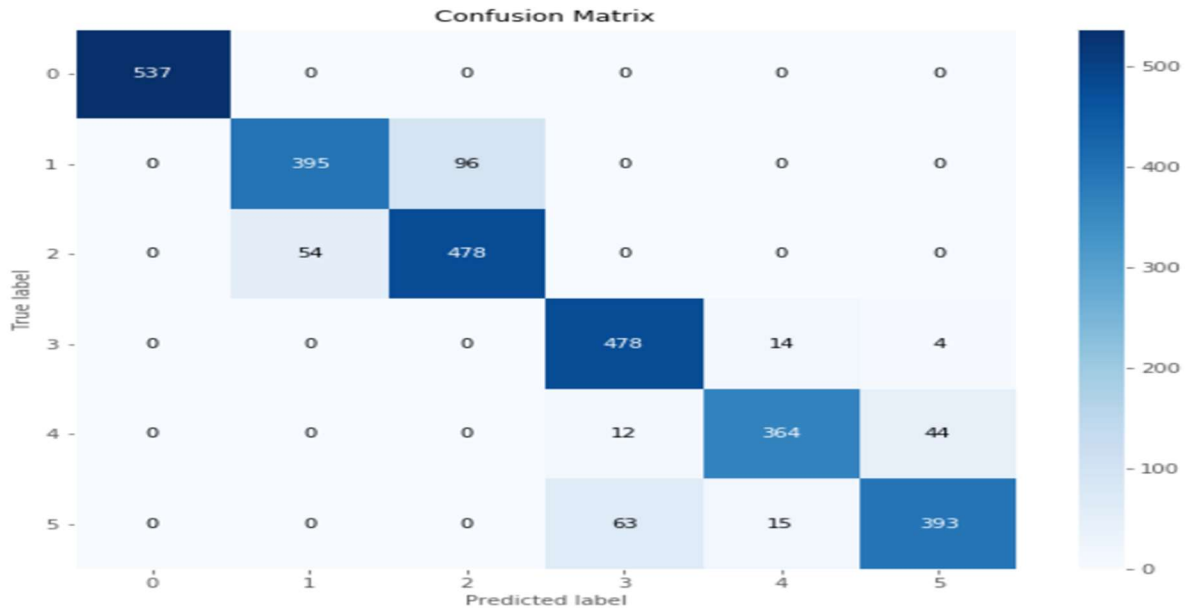


Fig 15: ANFIS confusion matrix

## ROC curve

ANFIS ROC curves for ANFIS model tested on dataset is shown in figure below:

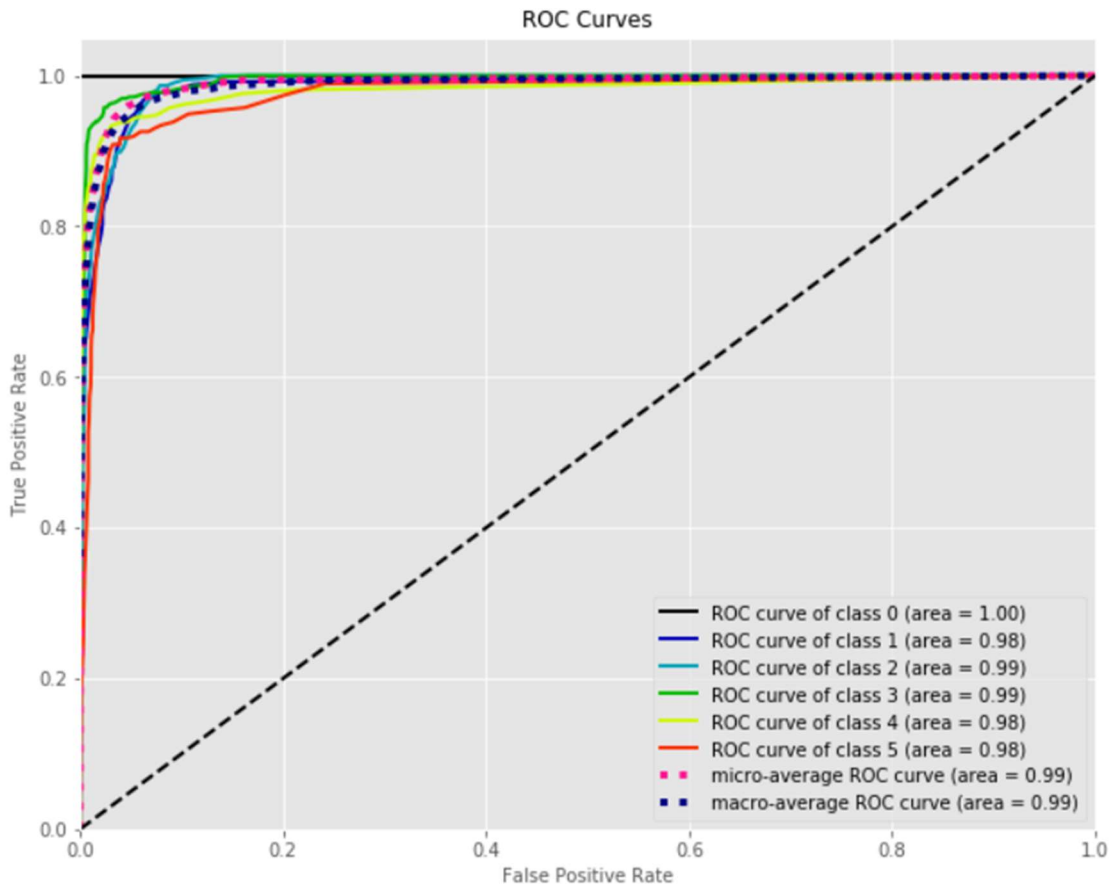


Fig 16: ANFIS ROC curve

### 4.4.2.2 Analysis of ANFIS algorithm

ANFIS classifier, in comparison to other classifiers, is not dependent on system description. ANFIS possess the advantage of the generation of the efficient model; even system parameters are unknown as in case of data from social media like tweets we use in this project. We begin ANFIS system training by generating fuzzy inference system using FCM clustering. The cluster generated by FCM clustering shows that cluster centroids for all the six labels are nearly identical to the desired value. The fuzzy inference system generated is tuned by using ANFIS algorithm. The training of the ANFIS system generates three linear parameters, twelve non-linear parameters, and twelve fuzzy rules. The membership function for the generated ANFIS system with input and output is shown in fig 16 below:

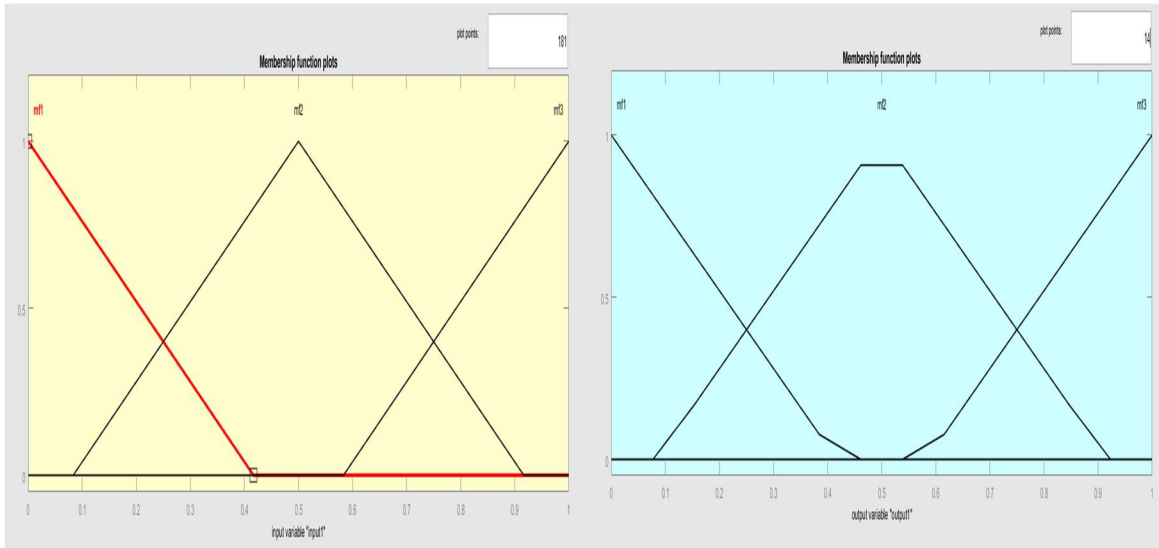


Fig 17: Input Membership function and output membership function

ANFIS system with membership function, rule base converts input data from crisp set to fuzzy values. ANFIS evaluates the rules, combines the output of each rule and finally converts the output data into the crisp form in an artificial neural network structure to generate a model for classification. We are using the evaluation metric the same as in case of k-mean, we can see that ANFIS performs with higher accuracy. The accuracy of the ANFIS system is 87.13% which is much better in comparison to the k-means algorithm. In our analysis, we can see that the use of ANFIS with FCM clustering generate a model that can classify tweets effectively in six labels.

Classifier	Accuracy
K-Means	73.41
Our Code	87.13

Table 6: Classifier accuracy

## **Chapter 5**

### **CONCLUSION AND FUTURE SCOPE**

The prime benefit of applying soft computing tools in our thesis instead of machine learning algorithm like the random forest is that we can build an effective classifier around the set of input-output value without requiring large no of features and mathematical model. We compared the accuracy of our proposed model in comparison to other well-known classifiers, and the result was satisfactory. One of the most deciding factors of good accuracy is the use of effective preprocessing, which, in our case sentiment score captures the sentiment of tweets. Sentiment score generation algorithm gives a comprehensive approach on how to incorporate unique characteristics of tweets as a whole. Emoticon based neuro-fuzzy system performs with increased accuracy and minimal execution time in in our project work when we compare it to other classifiers.

We are observing an accuracy of 87.14% with our system that does not mean we have solved sentiment analysis of social data. High accuracy means our system has certain limitations. Some of them are dense text preprocessing in the form of selecting only tweets having emoticons having clear information in emoticon ranking website, and the use of emoticon prevent substantial penalty of words that can change the meaning of the entire sentence. ANFIS system also poses a limitation of extensive iteration of selecting parameters and number of the epoch that affects the model performance.

To improve our system performance some of the work to be done in future scope are listed below:

1. Use of bigger dataset to improve system effectiveness.
2. Adding subjectivity of tweets in our sentiment generation methodology.
3. Strengthening of ANFIS model by combining with other evolutionary techniques like Genetic algorithm.
4. Integration of correlation of membership degree of classes in designing of fuzzy inference system

## References

- [1] R. Nagamanjula and A. Pethalakshmi, "A novel framework based on bi-objective optimization and LAN2FIS for Twitter sentiment analysis," *Soc. Netw. Anal. Min.*, vol. 10, no. 1, pp. 1–16, 2020.
- [2] M. Kanakaraj and R. M. R. Guddeti, "NLP based sentiment analysis on Twitter data using ensemble classifiers," *2015 3rd Int. Conf. Signal Process. Commun. Networking, ICSCN 2015*, 2015.
- [3] H. Wang and J. A. Castanon, "Sentiment expression via emoticons on social media," *Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015*, pp. 2404–2408, 2015.
- [4] Y. Yamamoto, T. Kumamoto, and A. Nadamoto, "Role of emoticons for multidimensional sentiment analysis of twitter," *ACM Int. Conf. Proceeding Ser.*, vol. 04-06-December-2014, pp. 107–115, 2014.
- [5] M. N. M. Salleh and K. Hussain, "A Review of Training Methods of ANFIS for Applications in Business and Economics," *Int. J. u- e- Serv. Sci. Technol.*, vol. 9, no. 7, pp. 165–172, 2016.
- [6] Z. Zhou, X. Zhang, M. Sanderson, H. Wang, and M. A. Sharaf, "Sentiment analysis on twitter through domain-specific lexicon expansion," *25th Australas. Database Conf. (ADC 2014) Databases theory Appl.*, pp. 98–109, 2014.
- [7] A. Al-Hmouz, J. Shen, R. Al-Hmouz, and J. Yan, "Modeling and simulation of an Adaptive Neuro-Fuzzy Inference System (ANFIS) for mobile learning," *IEEE Trans. Learn. Technol.*, vol. 5, no. 3, pp. 226–237, 2012.
- [8] E. Kolycheva Née Nikandrova and V. Kyrki, "Task-specific grasping of simiiar objects by probabiiistic fusion of vision and tactiie measurements," *IEEE-RAS Int. Conf. Humanoid Robot.*, vol. 2015-December, pp. 704–710, 2015.
- [9] S. M. Basha, Y. Zhenning, D. S. Rajput, N. C. S. N. Iyengar, and R. D. Caytiles, "Weighted fuzzy rule based sentiment prediction analysis on tweets," *Int. J. Grid Distrib. Comput.*, vol. 10, no. 6, pp. 41–54, 2017.
- [10] C. Jefferson, H. Liu, and M. Cocea, "Fuzzy approach for sentiment analysis," *IEEE Int. Conf. Fuzzy Syst.*, no. July, 2017.
- [11] C. Jefferson, H. Liu, and M. Cocea, "Fuzzy approach for sentiment analysis," *IEEE*



- Int. Conf. Fuzzy Syst.*, 2017.
- [12] G. Leroy and J. E. Endicott, “Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty,” *IHI’12 - Proc. 2nd ACM SIGHIT Int. Heal. Informatics Symp.*, pp. 749–753, 2012.
  - [13] M. Kanakaraj and R. M. R. Guddeti, “Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques,” *Proc. 2015 IEEE 9th Int. Conf. Semant. Comput. IEEE ICSC 2015*, pp. 169–170, 2015.
  - [14] R. Gaizauskas and K. Humphreys, “A Combined IR/NLP Approach to Question Answering Against Large Text Collections,” *Proc. 6th Content-Based Multimed. Inf. Access Conf. (RIAO-2000*, no. 1969, pp. 1288–1304, 2000.
  - [15] F. Ciravegna, “Adaptive information extraction from text by rule induction and generalisation,” *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 1251–1256, 2001.
  - [16] R. Gupta, P. Hooda, and N. K. Chikkara, “Natural Language Processing based Visual Question Answering Efficient : an EfficientDet Approach,” no. Iccics, pp. 900–904, 2020.
  - [17] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, “Sentiment of emojis,” *PLoS One*, vol. 10, no. 12, pp. 1–22, 2015.
  - [18] D. Kalbande, H. Panchal, N. Swaminathan, and P. Ramaraj, “ANFIS based Spam filtering model for Social Networking Websites,” *Int. J. Comput. Appl.*, vol. 44, no. 11, pp. 32–36, 2012.
  - [19] D. Karaboga and E. Kaya, “Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey,” *Artif. Intell. Rev.*, vol. 52, no. 4, pp. 2263–2293, 2019.
  - [20] M. Leizerovich, J. A. Martunova, ; Vladislav, and B. Sirotkin, “Neuro-fuzzy recruitment system Sistema de reclutamiento Neuro-Difuso,” *Páge*, vol. 38, 2017.
  - [21] M. K. Sohrabi and F. Hemmatian, “An efficient preprocessing method for supervised sentiment analysis by converting sentences to numerical vectors: a twitter case study,” *Multimed. Tools Appl.*, 2019.
  - [22] S. Yu, H. Zhu, S. Jiang, Y. Zhang, C. Xing, and H. Chen, “Emoticon analysis for Chinese social media and e-commerce: The azemo system,” *ACM Trans. Manag. Inf. Syst.*, vol. 9, no. 4, 2019.

- [23] A. Jayakumar, A. Shaji, and L. Nitha, “Wildfire forecast within the districts of Kerala using Fuzzy and ANFIS,” *Proc. 4th Int. Conf. Comput. Methodol. Commun. ICCMC 2020*, no. Iccmc, pp. 666–669, 2020.
- [24] A. Chavan, “Emoticon based Sentiment Analysis of Tweets.”
- [25] L. A. Ramos, A. F. Ramos, M. Melgarejo, and S. Vargas, “Tuning up Fuzzy Inference Systems by using optimization algorithms for the classification of solar flares,” *Tecciencia*, vol. 12, no. 23, pp. 35–46, 2017.