International Journal on Emerging Technologies **10**(3): 352-362(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Gurmukhi Text Emotion Classification System using TF-IDF and N-gram Feature Set Reduced using APSO

Ramandeep Kaur and Vijay Bhardwaj
Department of Computer Science and Engineering,
Guru Kashi University, Talwandi Sabo (Punjab), India.

(Corresponding author: Ramandeep Kaur)
(Received 26 June 2019, Revised 29 August 2019 Accepted 25 September 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Social media posts express people moods, emotions and present states of mind where people from different communities, origins, regions assemble together to share their views. Gurmukhi is Punjabi language based script which is widely used in Indian Punjab state. In this work, Emotion detection in Gurmukhi has been carried out on Gurmukhi dataset collected using Twitter API. Collected content has been filtered and six categories of data have been generated named as Sensitive, Happy, love, and Religious, Sad and Angry moods of people. There were 4237 documents left after filtering. For feature extraction, TF-IDF and N-gram features are used which are reduced further using MI (mutual information) and Particle swarm optimization (PSO) as larger number of Feature Set reduced classification speed of classifiers and the accuracy of classification improved. Most differentiable features are selected and most similar features are reduced in this step which results in high classification rates. Objective function used in feature selection by PSO is to maximize the area under curve (AUC) where Decision tree is applied as a classifier. Feature Sets having high AUC values are chosen and projected for classification by three classifiers named as Decision tree, Naïve-Bayes and k nearest neighbor. It has been analyzed that by applying TF-IDF word level and N-gram and feature reduction through MI-PSO, performance of emotion classification is above 90% by three classification procedures evaluated in positions of F-Score, Accuracy, Recall and Precision performance parameters. Among the classifiers, Decision tree founds most effective which has highest performance for three categories and in rest it almost gives high accuracy and F-measure rates.

Keywords: TF-IDF, N-gram, Gurmukhi script (Punjabi), PSO, AUC, feature reduction, Decision tree, Emotion classification etc.

I. INTRODUCTION

Sentiment analysis is used in almost every important field. It is used to depict people's mood, emotions and their reviews about various products [1]. A crucial role is played by emotions in our day to day activities affecting several phases of our existence comprising behavior, social interaction, decision-making and attitude [2]. How the individual sense and considering social emotion outlines plays vital role in numerous applications for instance public safety and health, urban planning and emergency response. Text is a predominantly significant basis of data for noticing emotion as the majority of word-based data extending from emails, micro blogs to SMS emails on a smart mobile which has turn out to be progressively accessible. The fast evolution of emotion-rich documented data creates a requirement to computerize analysis and identification of people's sentiment conveyed in text. Sentiment in societal networks and micro blogging tools (i.e., Facebook, Twitter) are progressively utilized by persons to share their feelings and opinions in way of tiny messages (e.g., text regarding normal lifetime and view on present events and problems) [3]. These messages (frequently well-known as micro blogs or tweets) can also comprise signs of sentiments of persons like anxiety, gladness, and unhappiness. Indeed, social webs comprise a huge amount of communal real-time data which is rich with sensitive content. It marks them suitable data bases for social revisions, particularly for reviewing sentiments of characters other than greater people.

Consequently, communal links like Twitter offer valued info to perceive mass reaction and conduct and learning a diversity of social characteristics and behavior [4]. Collective proof recommends that sentiment recognition and screening built about social mass media would be operative in numerous applications. Particularly, Twitter offers valued chances to detect community behavior and mood. The progress of vigorous word-based sentiment identifying tools assurances to have a considerable influence on individual and public well being and metropolitan planning. These sensation mining technologies, when accessible, can possibly be active in a huge range of applications extending from people level revisions of sentiments, the setting up of psychological health treatment amenities over social means, and further sentiment executive submissions. The census bureau and additional polling administrations can be capable to utilize the sentiment mining skill to approximate the proportion of individuals in communal undergoing definite emotions and relate this with existing actions and numerous other features of metropolitan living situations. This sort of skill could also improve primary outbreak warning for community health specialists with the purpose of a rapid act could happen. Furthermore, the sentiment mining technologies can also be utilized by advising organizations to screen emotional conditions of persons or to identify worry or complete stressors of people. On behalf of illustration, campus advising centers can be advised initially regarding upset scholars that might necessitate additional individual valuation. Challenges of detecting emotion in social networks have already been discussed [5]. English language has acquired a special place not only in our country but across the globe. It is necessary to have knowledge about this language to build a communication between different countries [6]. Punjabi language is an Indo-Aryan linguistic articulated by around 130 million publics residing in Punjab province of India and Pakistan. Punjabi is transcribed in two dissimilar scripts namely Shahmukhi and Gurmukhi.

Punjabi Language is transcribed in Gurmukhi script in India. The Gurmukhi script comprises 35 letters and the leading three letters are vowels. Six additional consonants are formed by retaining a dot at the base of the consonant and are utilized typically for loan words. In summary, the main contributions of this paper are:

(i) We created a Gurmukhi dataset for emotion detection from a large set of users using Twitter API. We used the data the show the efficiency of the proposed model in predicting multiple emotion states of users.

(ii) We proposed and implemented a machine learning model in which feature reduction phase has been introduced using PSO optimizer to exclude less relevant features and classification process has been carried out using reduced number of features.

II. RELATED WORK

In 2014, Bruno Trstenjak *et al*., (2014) implemented a structure for text classification founded on the TF-IDF method and KNN procedure. The foremost enthusiasm for investigation was to advance notion of contexts with importance on TF-IDF & KNN module. The outline with surrounded approaches provide upright consequences, established our notion and preliminary prospects. Assessment of outline was executed on numerous classes of forms in online setting. Trials are thought to offer responses regarding the excellence of cataloging and to regulate which issues have an influence on act of organization. The context exertion was reliable and stable. Throughout testing the superiority of classification they have attained decent outcomes irrespective of the K factor worth in the KNN procedure. Accomplished tests have noticed a sensitivity of the applied procedure. Tests revealed that the implanted procedure is profound to the sort of documents. The examination of documents matters displayed that the quantity of unusable words in documents has an important influence on the concluding superiority of classification. As, it is essential improve the preprocessing of data for attaining enhanced outcomes [7]. However, Castro *et al*., in 2017 offered a supervised technique built on n-gram model to categorize twitter data in both Portuguese nationwide variants: European from Portugal and Brazilian from Brazil. So as to seizure syntactic, orthographic, and lexical differences, these sorts of n-gram language representations were implemented: (1) word unigram and (2) character n-gram (from 'n' ranging from 2 to 7 characters) and (3) bigram. These sets of n-gram features were applied to construct ensemble and single models that were assessed by Logistic Regression, Naïve Bayes and two varieties of Support Vector Machine classifiers. They linked the leveled n-gram linguistic models in cooperation with the TF-IDF weighting system. Additionally, for the offered ensemble models, the class labels production were united by means of main stream algebraic and voting combiners like: minimum, mean, median, maximum, and product [8]. Furthermore, in 2017 Dey *et al*., build n-gram sentiment features by first mining the emotion words and their intensifiers from assessments. The totals equivalent to these features are attained from the present emotion lexicons. Proposed Lexical TF-IDF matrix is created by multiplying TF-IDF rating with feature score. They research with two benchmark data sets and two renowned classifiers with cross domain authentication demonstrates that their method outdoes in 81.25% cases allowing for all the act actions, therefore, could be applied for real data sets wherever sample designs are not accessible [9]. None the less, in 2018, Bansal and Srivastava use word2vec demonstrations to categorize above 400,000 online customer evaluations for numerous intercontinental mobile phone products attained from Amazon. They first novelty features utmost analogous to manufactured goods features by word2vec and demonstrate that word2vec is capable to discover semantically analogous words. At that time they applied skip-gram and CBOW approaches with four dissimilar classification procedures: Naïve Bayes, SVM, Random Forest and Logistic regression. Outcomes demonstrate that CBOW executes fine in contrast to skip-gram, represent that data might comprise commonly happening identical words. Random Forest outstrips all the procedures once applied with word2vec demonstrations. Here after, distributed word vector demonstrations could be proficiently active for the undertaking of sentiment ordering by integrating semantic word associations and relative information [10]. and Laith Mohammad Abualigah *et al*., offered a novel scheme to resolve the text feature selection problematic applying an unsupervised learning process. Particle swarm optimization (PSO) is utilized as a feature selection procedure. This procedure utilizes the term frequency-inverse document frequency (TF-IDF) as an impartial function to assess each text feature at the level of the document. The proposed method (FAPSOTC) considers the innovative dataset to attain a novel optimal sub-set of instructive features. The novel subset of informative features is passed in to k-mean text grouping procedure to explore the feature selection technique conferring to the group accuracy [11]. At this point in time, Ashima Kukkar *et al*., applied the hybrid methodology of integrating natural language processing, text mining and machine learning procedures to categorize which bug crash as bug or non-bug. Because of this the clamors of mis-classification is abridged by sieving the bug reports and boost the act of instinctive bug guess. In this work the four integrate fields (severity, reporter, component and priority) with textual data such as description; comments, summary etc are additional to training and testing dataset. The TF-IDF and bigram approach is applied with information improvement for the fault severity estimation. The bigram methodology assisted in decreasing the sparisity of dataset. So as to compute the accurateness of offered model, Recall, Precision and F-measure are utilized [12]. And Mohammad Razzaghnoori *et al*., presented certain approaches for query arrangement founded on Word2vec vector demonstration that could seizure a huge data of accurate semantic and syntactic associations. They detected that the consequence of tf-idf weighting on refining the accurateness is note worthy; nonetheless the tf-idf factor must be judiciously altered to obtain improved consequences. They have extended reasonable consequences seeing the information that these queries were considered to be tough to response equal for persons [13]. Furthermore, Tuba Parlar *et al*., (2018) explored the possessions of four term weighting approaches on the sentimentality examination of Turkish evaluations. Additionally, they inspected the communications among feature selection and term weighting approaches for text depiction utilizing NBM classifier on five Turkish assessment datasets. The investigational consequences illustrate that dissimilar term weighting approaches do mark the act for the sentimentality examination of Turkish evaluations and tf* idf weighting technique provides the greatest F degree values for the concentrated text illustration attained over feature selection.

Table 1: Tabular Survey.

Ref. No.	Title	Feature Extraction/Selection	Classifier	
[10]	Emotion classification of online consumer assessments using word vector representations	skip-gram methods	SVM, Naïve Bayes, Random Forest and Logistic regression	Amazon review
[7]	KNN with TF-IDF Based Framework for Text Classification	TF-IDF	KNN	Sets of 500 online documents from different categories
[11]	A novel feature selection technique to advance the document clustering by applying particle swarm optimization procedure	Particle swarm optimization (PSO)	k-mean text clustering procedure	Six text document datasets from Reuters–21578 and 20Newsgroups.
[12]	A Supervised Bug Report Organization with Incorporate and Textual field Knowledge	TF-IDF and Bigram	K-nearest neighbor (K-NN)	Five open source project (Eclipse, Mozilla(core), Firefox, J Boss, Open FOAM)
[14]	Query organization in Persian using word vectors and frequencies	Continuous bag of words and continuous skip-gram models	Neural Network classifiers and Support Vector Machine	UTQD.2016 contains 1175 Persian questions
[8]	Smoothed n-gram based models for tweet language identification: Acase study of the Brazilian and European Portuguese national varieties	TF-IDF weighting	Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression, and Log-Likelihood Ratio(LLR)	385, 710 tweets
[9]	Lexical TF-IDF: An n-gram Feature Space for Cross-Domain Classification of Sentiment Reviews	N-gram	Maximum Entropy method and Support Vector Machine	IMDb (2004) and Epinion
[14]	Interactions among Feature Selection and Term Weighting Methods on the Sentiment Analysis of Turkish Reviews	IDF Cutoff, χ^2 and DFD	NBM classifier	five Turkish review datasets
[15]	A novel feature extraction organization for emotion examination of product evaluations	G_TF–IDF, FPCD, T_TF–IDF+FPCD, G_TF–IDF+FPCD	KNN, NB, RF, LR, SVM, DT, GBDT	ChnSentiCorp-Htl-del-4000, ChnSentiCorp-Nb-del-4000, IMDB review
[16]	Question classification with log-linear models	Lexical and syntactic information: POS tags, language model, chunks, Word Net for target, named entity tags, and quoted strings	ME (maximum entropy model)	UIUC
[17]	A semantic methodology for question classification utilizing WordNet and Wikipedia	Words, named entity, semantic information		UIUC
[18]	Question classification using head hyponyms and words	Head word, wh-words, Ngram, WordNet semantic feature for head word, word shape	ME and SVM	UIUC
[19]	Question organization in Persian language based on provisional random fields	words, Question informer, N-gram, Question Words, position of tokens, POS tags	CRF	Almost Primary and junior high school
[20]	A hybrid methodology for question organization in Persian automatic question answering systems	Lemm, POS tags, length of question, n-gram, normalized word, special word detection, Verse Finder	SVM with defined rules	Quranic Question
[21]	Latent semantic examination for question organization with neural networks	word-shapes, Bigrams, wh-words, related words, headwords, hyponyms	SVM and BPNN	UIUC

Additionally, together χ^2 and DFD support create the finest consequences for certain datasets, though χ^2 inclines to effort fine with slighter feature sizes, whereas DFD inclines to favor superior feature sizes. They also examined these weighting approaches utilizing NBM classifier on the English analysis datasets. Contrast with Turkish assessments, tp mechanism the finest as the baselines and χ^2 is the top performer for all five English datasets. Both tp and tf* idf approaches effort rationally well for these datasets with minor alterations in the act, signifying that with a compact number of features, tf* idf could be operative in discerning amongst the particular features. For English as well as Turkish assessment datasets, the IDF cut off technique hold ups overdue the added two feature selection approaches, nevertheless associated through the baseline consequences, it could attain improved classification

consequences with smaller feature sizes, ranging among 1500 and 3500 features. Generally, the enhancements are altogether noteworthy once conjoining feature selection and term weighting approaches, demonstrating that this grouping is important and essential for the emotion examination of English and Turkish reviews [14]. Additionally, Xin Chen et al planned a new feature extraction approach for Sentiment Analysis (SA) of Chinese evaluations. Two sorts of features are controlled in the procedure: one is the OPSM bi-clustering emotion feature of comprehensive TF–IDF and the additional is the recognized FPCD phrase feature identified built on the enhanced Prefix-Span procedure. The previous can lessen the sparsity over synonym computation based on Word2vec. Conversely, they create complete usage of the comparative size of occurrences of words incidence that evades the problematic of adaptable text length. The Prefix-Span is enhanced with branch-and bound approach and so the conforming FPCD phrase feature comprises the word-order info in the assessment and confines the recurrent phrase designs by the extreme gap and smallest discriminative capability threshold of words. So the FPCD phrase has durable discriminative capability and additional progresses classification accuracy [15].

III. SYSTEM MODULE

Presented system module includes four phases (i) Pre-processing (ii) Feature extraction (iii) Feature selection and (iv) Classification. Pre-processing involves filtering of data in which stop-words, special characters are removed and a 3-Gram corpus is designed for feature extraction process. Further feature extraction is carried out in which TF_IDF has been evaluated for the generated corpus which is further concatenated for feature selection process. As the dimension of composed feature matrix is high, feature selection phase is introduced by applying PSO. Before feeding feature set into PSO, feature sorting process is carried out that assembles the features according to differentiability among them. Finally classification process is carried out in which NB, kNN and decision tree are used for validation of proposed method. The flow chart of the proposed scheme is given in Fig. 1.

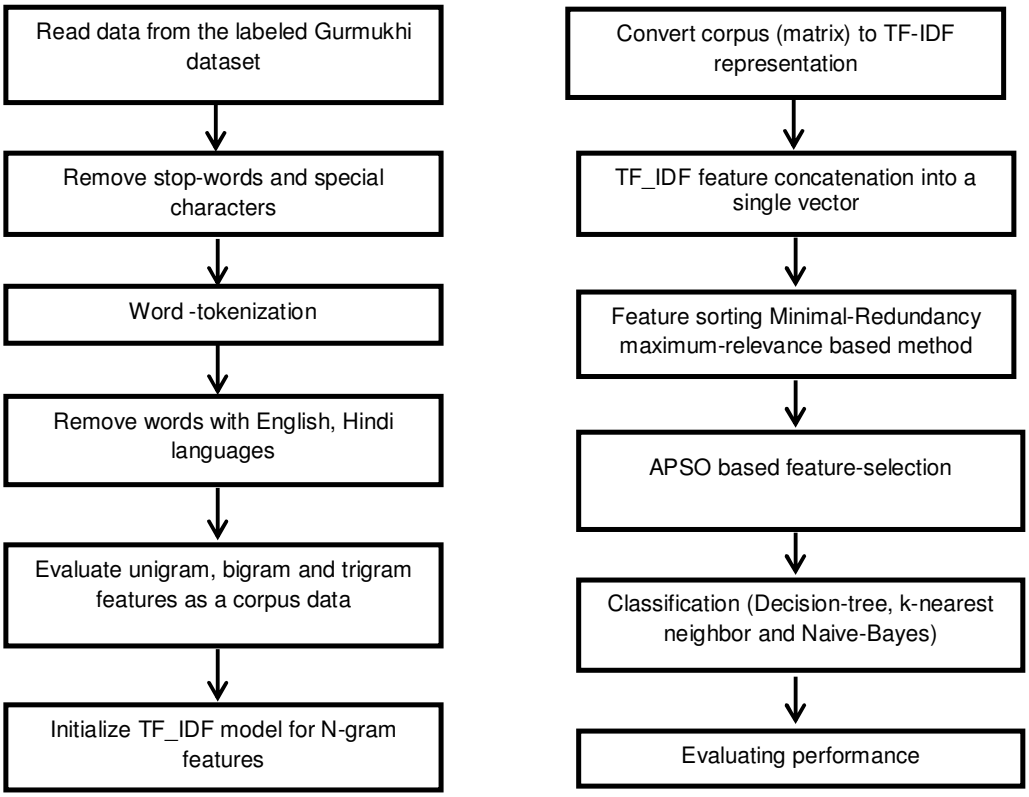

Fig. 1. Flowchart of the Methodology adopted.

A. Feature Extraction

1. Term Frequency-inverse Document Frequency (TF-IDF)

In order to evaluate the significance of a word in a document TF-IDF is used which is a well-defined method widely used in sentiment analysis and text categorization. TF of a particular term (t) is find out as a no. of time that particular term occurs in a review or a document with respect to total terms in the document. Inverse document frequency (IDF) is used to evaluate the significance of the word. Inverse document frequency is evaluated as IDF (t) = log (N/DF) where N is number of documents and DF is number of documents which has that term. IF_IDF is effective in converting the textual information in a vector-space model.

Suppose a document has 300 words and scissor appears 40 times in these 300 words. Than term-frequency (tf) = 40/300 = 0.1333 and let out of total 40000 documents, there are 100 documents which contains scissor word. Than IDF (scissor) = 40000/100=400, and TF-IDF (scissor) will be 0.1333*400= 53.32

$$\text{weight}_{i,j} = \text{term-frequency}_{i,j} \times \log \frac{\text{Number\_docs}}{\text{document-frequency}_i} \quad (1)$$

In a study completed in 2017, it was concluded that term weighting technique like tf-idf accomplished well for feature selection in circumstances of high dimension feature space [22].

2. N-gram (Phrases)

N-gram is also used as a feature extraction method when machine learning classifiers are used for text classification process [23]. These comprises of n number of tokens which comes in sequence in a given text. N can be any number 1, 2, 3 and so on but mostly first three unigram, bigram and trigram are widely used. If we point out a sentence "Artificial intelligence is the future" and consider N = 2 than it will produce "Artificial intelligence", "intelligence is", "is the", "the future".

B. Feature matrix sorting using Minimal redundancy maximal relevance (MRMR)

To maximize the feature-relevancy and to minimize redundant features, a filter based approach has been introduced in [24] which use mutual-information (MI) in between discrete or continuous feature set. It uses an optimal first-order incremental selection to generate a candidate list of features that cover a wider spectrum of characteristic features. It converts the continuous features into discrete features first and can uses different levels of quantization) by localizing the limitations at μ ± σ that are predictable standard and mean-deviation correspondingly. These consequences in a set of discrete feature set y (f), f ∈ F. Method can be analyzed using the subsequent rule.

$$S_d = S_{d-1} \cup \arg\max \text{f}[\text{I}(\text{x(f)}, \text{class\_label}_z)$$

$$-\frac{1}{d-1}\sum_{g \in S_{d-1}} \text{Mutual\_Info}(\text{x}(f), \text{x}(g)] \quad (2)$$

The statistical dependency in between discrete features x and the category names is measured by the eq.

$$\text{standard\_deviation}(SD) = \sum_{x \in Y}\sum_{z \in Z} p(\text{x}, z)\frac{p(x,z)}{p(\text{x})p(z)} \quad (3)$$

High value of SD means higher relevancy in between the category and component features. If these components need to be independent from the category labels, the value of SD should be less than one. This degree of comparison can be carried out using the mutual information (MI) attribute as described in equation below.

$$\text{mutual\_information (MI)} = \sum_{x \in Y}\sum_{z \in Z} p(\text{x}, z)\log\frac{p(\text{x}, z)}{p(\text{x})p(z)} \quad (4)$$

C. Feature selection using Particle swarm Optimization

This [25] presented the idea of Particle Swarm Optimization (PSO) that originates from bird-flocking. Consider a group of words in a n-dimensional space searching for food and in start nobody has the food. They find the position of that word which is nearest to the food. By this way the rest of the words track the best nearest to the food. PSO describes each word as a particle whose location can be taken as assumption as given in equation below

$$x_j = (\text{x}_{j1}, \text{x}_{j2}, \text{x}_{j3} \ldots \text{x}_{jn})$$

This solution try's to converge to optimum solution after each iteration with a velocity of the particles assumed in start as

$$v_j = (\text{v}_{j1}, \text{v}_{j2}, \text{v}_{j3} \ldots \text{v}_{jn})$$

It gives the direction of the particle movement in the coming iteration, It can have +ve as well as –ve values. Local best defines the optimal best in the current iteration whereas global best defines the optimal best among all iteration reached till now. The novel rationalized velocity is provided as below:

$$\text{velocity}_{d+1} = k * (\text{weight} * \text{velocity}_d + \phi_1.r$$

$$\text{and } ()*(p_{\text{best}} - x_d) + \phi_2.r \text{ and } ()*(g_{\text{best}} - x_d)) \quad (5)$$

$$\text{velocity}_{d+1} = x_d + \text{velocity}_{d+1} \quad (6)$$

where ϕ_1 and ϕ_2 are acceleration factors, weight is the inertia-weight factor, k is the compression factor, r and () gives random values between 0 and 1. Acceleration factor resolves the group size of the particle in the upcoming iteration. The flow chart of the PSO procedure is represented in Fig. 2.

D. PSO and diversity measures for weight adaptation

Firstly, uninterrupted non-linear roles were resolved by utilizing PSO. The indication originates from bird-flocking. The variety restrictions in PSO shows main role in conjunction to optimum solution. Diversity measure means that a huge penetrating space is discovered by particles having lesser resources which reduce area investigated by the particles. It is moderately noteworthy to calculate the examiner conduct of a PSO procedure once swarm diversity is measured in the exploration as in the Repulsive and Attractive PSO [26]. Hence these procedures essentials to correct computing of search behavior in swarm with respect of one time to another time. The diversity measures and variables related with this are summarized below

1. The swarm diameter. It is the maximum distance in two particles in swarm [27].

$$|\text{Diameter}_{Dim}| = \underset{(i \neq j) \in 1,|S|}{\text{maximize}}\left(\sqrt{\sum_{k=1}^{l}(x_{ik} - x_{jk})^2}\right) \quad (7)$$

Where I and are the dimensionality and solution of the problem with swarm size |S| and x_{ik} is the k^{th} dimension of postion of the i^{th} particle.

2. The average-distance around center of swar

This formula is [25]:

$$\text{Distance}_s = \frac{1}{|S|}\sum_{i=1}^{l}\sqrt{\sum_{k=1}^{l}(x_{ik} - x_{jk})^2} \quad (8)$$

3. Swarm Coherence

$$\text{swarm coherence}(S_c) = \frac{V_s}{\bar{V}} \quad (9)$$

Where V_s= speed of the swarm center given as:

$$V_S = \frac{1}{|S|}||\sum_{i=1}^{|S|}\bar{v}_i||_2 \quad (10)$$

and $\bar{v}_i$ = average particle-speed of the

$$\bar{v} = \frac{1}{|S|}\sum_{i=1}^{|S|}||\bar{v}_i||_2 \quad (11)$$

Eqn. 7-11 is utilized as a degree of diversity in recommended APSO (Adaptive Particle swarm optimization) technique to calculate the weight for the subsequent iteration. In order to select the best particle in iteration, objective function is required based on which a fitness value is calculated for each particle. In classification based optimization, mainly Area Under Curve (AUC) is used as objective function which is evaluated for the selected features by PSO depending upon the number of classes or categories in the dataset that need to be classified.

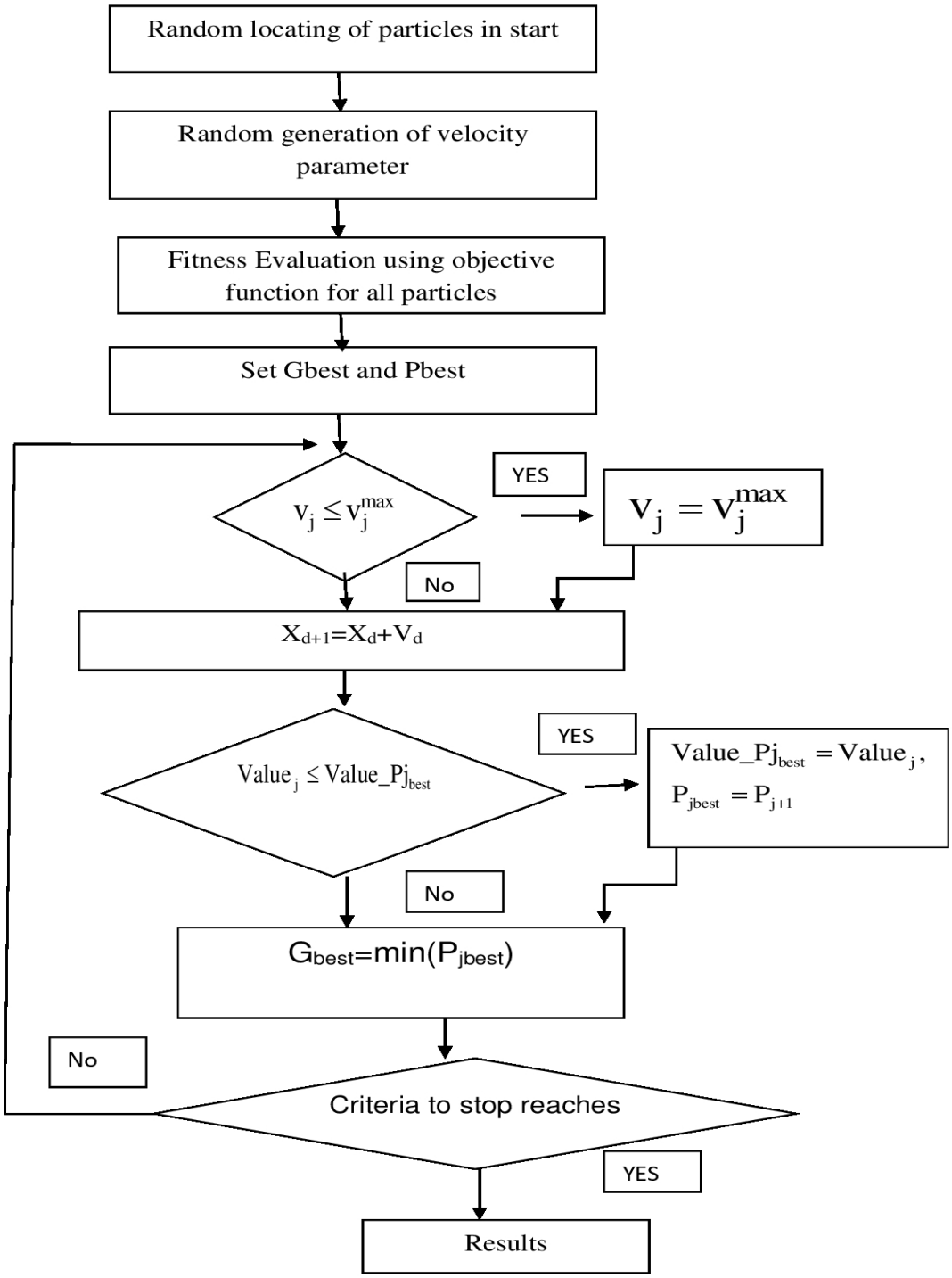

Fig. 2. Flowchart of proposed method.

Sometimes precision or sensitivity measures are also used and optimum best is figured out from the particles. The objective function applied is assumed in form of equation as below

$$\text{Objective function} = \text{maximize}\,(\text{mean}(\sum_{i=1}^{N} AUC)) \quad (12)$$

where N is the number of emotion categories

Hence forth the basic difference of adaptive PSO from the traditional PSO is insertion of inertia weight during coming iterations. The formula for new weights is given as:

$$w(t) = w(t-1) \times \left(\frac{1}{\text{Diameter}_{Dim} + \text{Distance}_S} + \text{alpha} \times S_v \right) \quad (13)$$

where w(t-1) is previous iteration inertia weight and weight (t) is current iteration inertia weight, alpha is a constant.

E. Feature Training step for text categorization

The final stage of text categorization process involves the training and testing the dataset using the selected feature set given by the feature selection phase using Adaptive particle swarm optimization. Three machine learning classifiers named as DT, NB, and k-NN are used for classification. As there is need in fitness function evaluation in APSO, decision tree is used to get the area under curve in which the chosen class was marked as true class whereas rest are marked as false. Similar procedure is adopted for rest of the classes after which mean of AUC is used as objective function and maximum value is chosen as the local best.

1. Naive Bayes (NB)

Bayesian theory can be used to text classification which allocates the class c = Argmaxc P (c|d), to assumed document d. The relation 14 is stated founded on the Bayesian theory.

$$P(c \mid d) = \frac{p(c)p(d \mid c)}{p(\mathrm{d})} \tag{14}$$

where p(d) has no role in selecting c. To approximate the term p (d|c), Naïve Bayes molds it by supposing the featureset fi's are provisionally autonomous stated d's class as in relation 15.

$$P_{NB}(c \mid d) = \frac{P(C)(\pi_{i=1}^{m} P(f_i \mid c)^{n_i(d)})}{p(d)} \tag{15}$$

where m is the number of features and f is the feature vector.

2. K-nearest neighbor (KNN)

KNN is one of the common examples of learning-based techniques. In this technique, K is the number of measured neighbors that is frequently odd, and the distance to such neighbors is known based on the Euclidean distance values. The fundamental supposition in this technique is that, each sample is real points in n-dimension space. Generally, this procedure is applied for two reasons: to approximate the distribution density of the used training data set and to categorize testing dataset depending upon the training dataset [28].

3. Decision tree (DT)

The decision tree [29] is one of the utmost well-known machine learning systems in which its primary objective is to estimate the objective tasks with discrete values. This tree is called as decision tree since it represents the method of decision making to regulate an input sample group. The decision tree could be a good selection for opinion mining since it has very decent performance in contradiction of the high-volume data.

IV. RESULTS AND DISCUSSION

Attaining a dataset for the sentiment documentation undertaking is additional stimulating as there are not any by hand categorized openly accessible datasets. This [1] describe 7 discrete human sentiments (joy, love, anger, surprise, fear, sadness, thankfulness). All of these sentiments they have induced set of keywords and their lexical variations to signify a single human sentiment. From these, we utilize joy (happy), love, sadness and anger alongside with religious and sensitive categories. The aim after this is the sort of Gurmukhi content current on twitters as many people post religious and emotional posts on twitter. At that time, the Twitter API was inquired for tweets comprising any of the keywords in the method of a hash tags. Furthermore, the quality of the dataset was manually assessed by arbitrarily sampling a minor share of the dataset for manual examination by human annotators. A text amount from Twitter is advanced in the existing revision by assembling thousands of posts spread on six classes. A total of 4237 tweets have been composed from six classes. For training purposes 70% documents are used for training in both evaluating the objective function for PSO and for final classification by different classifiers.

Table 2: Table showing percentage of samples classified accurately to class.

Emotion Class	Number of documents
Sensitive	507
Happy	1687
Love	377
Religious	284
Sad	712
Angry	670

The classification of view could be assessed utilizing four indexes considered on the base of the subsequent calculations: precision, accuracy, f-measure and recall [27]. Many documents detected and the ground truth category provides the effectiveness of the presented method based on sensitivity of particular class and specificity of rest of the documents. Hence Detection Accuracy (DA) or recall value which defines sensitivity parameter can effectively represent the accuracy of text classifier. Along with recall, precision, F-measure and Accuracy parameters are also used. The formulas

$$\text{Precision} \quad \frac{TP}{(TP + FP)} \tag{16}$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{17}$$

$$F\_\text{Measurement} = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{18}$$

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{19}$$

Where true positive is suitably categorized documents in a category, False-positive as appropriately classified documents in a category and false negative as correctly classified documents of rest categories; Gurmukhi emotion classification results for tested documents are represented in Table 3.

Proportion of appropriately expected positive observations to all observations in definite class yes is identified as recall. Decision tree gives high recall ratio for almost all categories which falls in range 0.82 to 1 except the happy emotion category where Naive-Bayes gives high recall value of one. For rest of the categories Naïve-Bayes gives lower recall value which falls in range 0.75 to 0.82. kNN gives moderate values which falls in between the recall values of both decision tree and Naïve-Bayes classifier.

Table 3: Performance evaluation.

Parameters	TP	TN	FP	FN	Precision	Recall	F-Measure	Accuracy
Emotion Category	**Classification Results Using Decision Tree**							
Sensitive	415	3739	0	92	1	0.8185	0.9002	0.9783
Happy	1609	2458	92	87	0.9459	0.9487	0.9473	0.9578
Love	327	3782	87	50	0.7898	0.8673	0.8268	0.9677
Religious	257	3912	50	27	0.8371	0.9049	0.8697	0.9818
Sad	646	3507	27	66	0.9598	0.9073	0.9328	0.9780
Angry	670	3510	66	0	0.9103	1	0.9530	0.9844
	KNN results							
Sensitive	412	3729	10	95	0.9763	0.8126	0.8869	0.9752
Happy	1594	2426	124	102	0.9278	0.9398	0.9338	0.9467
Love	348	3677	192	29	0.6444	0.9230	0.7589	0.9479
Religious	230	3949	13	54	0.9465	0.8098	0.8728	0.9842
Sad	583	3494	40	129	0.9357	0.8188	0.8734	0.9601
Angry	670	3546	30	0	0.9571	1	0.9781	0.9929
	Naïve Bayes results							
Sensitive	383	3739	0	124	1	0.7554	0.8606	0.9707
Happy	1696	2061	489	0	0.7762	1	0.8740	0.8848
Love	294	3869	0	83	1	0.7798	0.8763	0.9801
Religious	234	3952	10	50	0.9590	0.8239	0.8863	0.9858
Sad	556	3532	2	156	0.9964	0.7808	0.8755	0.9627
Angry	538	3532	44	132	0.9243	0.8029	0.8594	0.9585

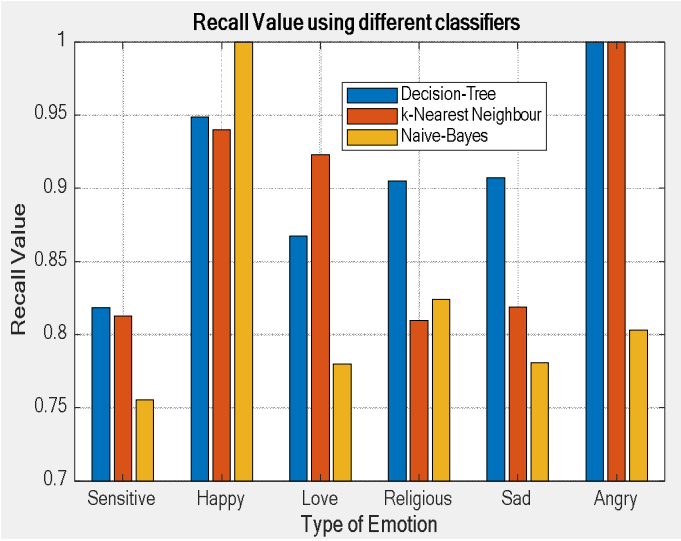

Fig. 3. Bar graphs for Recall Value for Gurmukhi Text emotion classification.

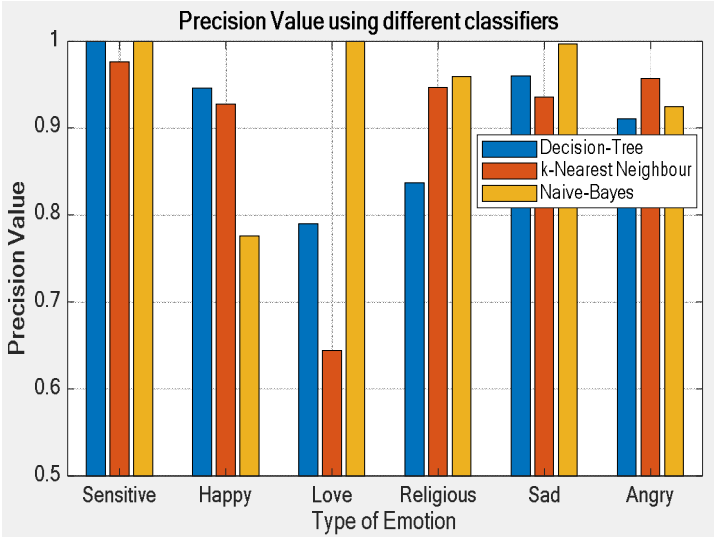

Fig. 4. Bar graphs for Precision Value for Gurmukhi Text emotion classification.

Proportion of expected positive observations to the whole positive observation is identified as precision. Precision value falls in between 0.7898 and 1 for decision tree, 0.6444 to 0.9571 for kNN where lowest precision is detected for Love emotion category and highest for angry category. Naïve-Bayes gives precision value in between 0.7554 and 1 with least precision value noted for sad category and highest for love category.

Weighted average of precision and recall is called f-score. It is additional significant limit than accuracy once having an uneven class distribution in data. Decision tree gives highest F-score value for three categories which is 0.9002, 0.9473 and 0.9328 for sensitive, happy and sad categories. F-measure value for Angry category is also effective which is 0.9530 but is less than kNN classifier which gives 0.9781 for angry category.

Naïve Bayes Has F-measure in between 0.8594 and 0.8863 in which it gives highest F-score for Love category when compared with other categories.

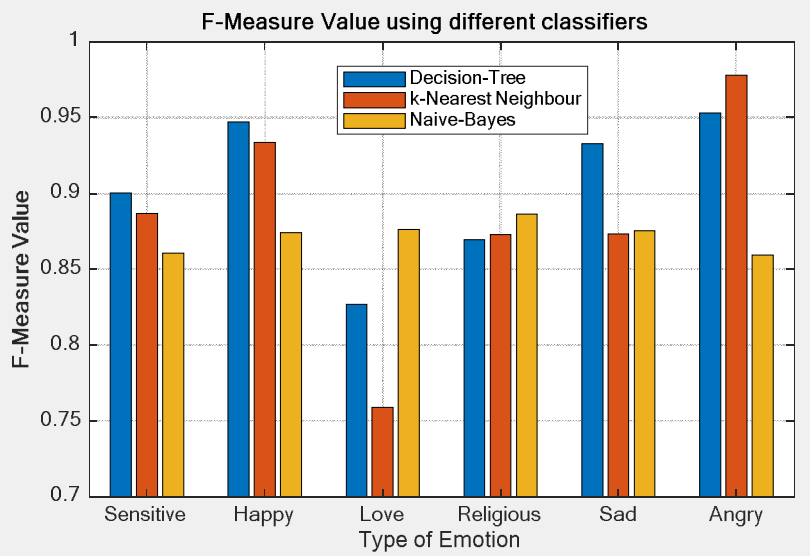

Fig. 5. Bar graphs for F-Measure Value for Gurmukhi Text emotion classification.

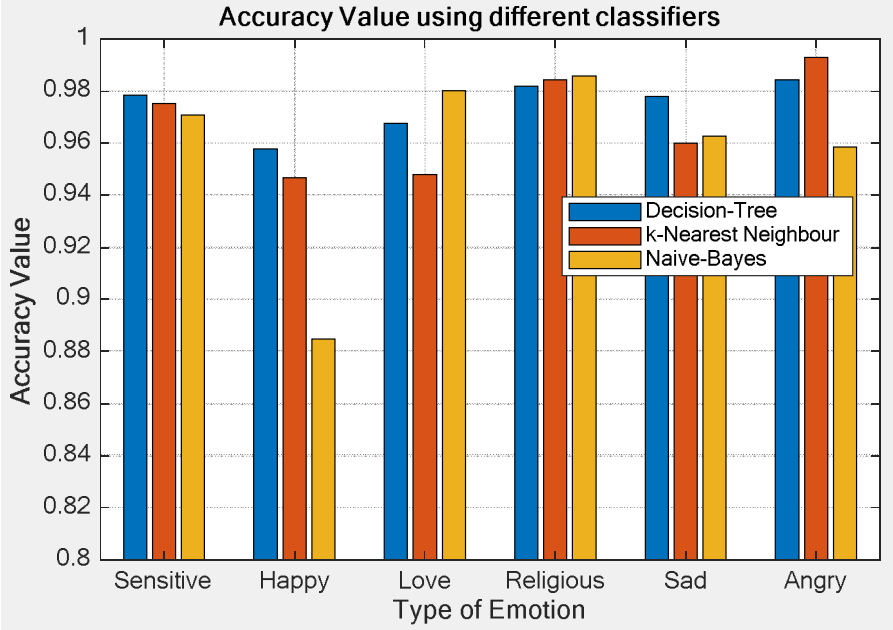

Fig. 6. Bar graphs for Accuracy Value for Gurmukhi Text emotion classification.

This is the proportion of true negative plus true positives to the true negatives plus true positives plus false negative plus false positive as presented in Eqn. 1. It computes how much measurement of cases is appropriately classified. Decision tree gives highest accuracy among all three tested classifiers in which DT gives highest classification accuracy for three categories and for rest three categories it is almost nearest to maximum performance by the others. It gives 0.9783, 0.9578, 0.9677, 0.9818, 0.9780 and 0.9844 accuracy values for sensitive, happy, love, religious, sad and angry categories respectively. Naïve-Bayes gives highest accuracy for love (0.9801) and religious (0.9858) emotion categories.

V. CONCLUSION

In this paper, two features TF-IDF (word level) and N-Grams (value of n = 2 and n = 3) are considered to carry out emotion classification on Twitter dataset collected for Punjabi language with Gurmukhi script. Dataset has been collected using twitter API in which Gurmukhi script filtering has been applied which eliminates Hindi, English or other language data and provides only Gurmukhi content from the tweets. Further this data has been filtered and only content having six emotion categories named as sensitive, happy, love, religious, sad and angry has been considered. Classification results are shown for four presentation parameters i.e. precision, accuracy, f-score and recall in which three classification methods Naive Bayes, Decision Tree and KNN are used for classification. Proposed method constitutes four steps (i) Preprocessing in which filtering of stop words etc. has been carried out (ii) feature extraction in which unigram, bigram and trigram data corpus is generated and further TF-IDF features are produced (iii) feature selection in which MI (Mutual Information) and PSO procedure are used for selecting less but effective feature set iv) classification in which three classifiers are used for evaluating category of the emotions.AUC is used as objective function in PSO procedure in which selected feature set by PSO is used in classification process and AUC is evaluated for each category. Finally mean of AUC is evaluated and Best solution is found in selecting feature set that has maximum mean AUC value. Experimental results shows that decision tree is the best which shows high values of recall, F-score and accuracy parameter for most of the categories which falls in range 0.82 to 1. This system is used to help companies to obtain qualitative insights to understand how people are talking about their brand.

VI. FUTURE SCOPE

Sometimes people prefer to visualize their textual content in order to make posts attractive and upload relative image to the text. Future research will be continued to develop multi-modal emotion detection system to such posts in which images as well as text will be used and weight based classification system will be generated for both images and text.

ACKNOWLEDGEMENTS

I would like to express here the very thanks to my supervisor Dr Vijay Bhardwaj. This paper would have been not possible without his immense guidance and inspiration. He has provided me extensive personal and professional knowledge and taught me a great deal about scientific research. Nobody has been more important to me in the pursuit of this paper than the members of my family. I would like to thank my parents. They are the ultimate role models. Most importantly, I wish to thank my supportive husband Gurniwaz Singh Sandhu, who provided an unending inspiration.

REFERENCES

[1]. Venkatesan, D., Ruba, K. V., & Sekar, K. R. (2019). Investigation of various Techniques and Classification Methods on Cognitive Sentimental Learning. *International Journal on Emerging Technologies, 10* (2), 15-18.

[2]. Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing Twitter 'Big Data' for Automatic. *In proceedings of International Conference on Social Computing (Social Com)*, 587–592.

[3]. Choudhury, M. D., Counts, S., & Gamon, M. (2012). Not all moods are created equal. Exploring Human Emotional States in Social Media. *In proceeding of Sixth international AAAI conference on weblogs and social media.* Redmond: ICWSM.

[4]. Wakamiya, S., Belouaer, L., Brosset, D., Lee, R., Kawai, Y., Sumiya, K., & Claramunt, C. (2015). Measuring crowd mood in city space through twitter. In *International Symposium on Web and Wireless Geographical Information Systems* (pp. 37-49). Springer, Cham.

[5]. Hasan, M., Rundensteiner, E., & Agu, E. (2018). Automatic emotion detection in text streams by analyzing Twitter data. *International Journal of Data Science and Analytics*, *7*(1), 35-51. doi:10.1007/s41060-018-0096-z

[6]. Upadhyay, O. (2018). A Comparative Study of English for Specific Purpose (ESP) and English as a Second Language (ESL) Program for hote management Students. *International Journal on Arts, Management and Humanities , 7*(1), 27-32.

[7]. Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based Framework for Text Categorization. Procedia Engineering, 69, 1356-1364. doi:10.1016/j.proeng.2014.03.129

[8]. Castro, D., Souza, E., Vitorio, D., Santos, D., & Oliveira, A. L. (2017). Smoothed n-gram based models for tweet language identification: A case study of the Brazilian and European Portuguese national varieties. *Appl. Soft Comput., 61*, 1160--1172.

[9]. Dey, A., Jenamani, M., & Thakkar, J. J. (2017). Lexical TF-IDF: An n-gram Feature Space for Cross-Domain Classification of Sentiment Reviews. Lecture Notes in Computer Science Pattern Recognition and Machine Intelligence, 380-386. doi:10.1007/978-3-319-69900-4_48

[10]. Bansal, B., & Srivastava, S. (2018). Sentiment classification of online consumer reviews using word vector representations. Procedia Computer Science, 132, 1147-1153. doi:10.1016/j.procs.2018.05.029

[11]. Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *Journal of Computational Science*, *25*, 456-466. doi:10.1016/j.jocs.2017.07.018.

[12]. Kukkar, A., & Mohana, R. (2018). A Supervised Bug Report Classification with Incorporate and Textual field Knowledge. *Procedia Computer Science*, *132*, 352-361. doi:10.1016/j.procs.2018.05.194.

[13]. Razzaghnoori, M., Sajedi, H., & Jazani, I. K. (2018). Question classification in Persian using word vectors and frequencies. *Cognitive Systems Research*, *47*, 16-27. doi:10.1016/j.cogsys.2017.07.002

[14]. Parlar, T., Özel, S. A., & Song, F. (2018). Interactions Between Term Weighting and Feature Selection Methods on the Sentiment Analysis of Turkish Reviews. Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, 335-346. doi:10.1007/978-3-319-75487-1_26

[15]. Chen, X., Xue, Y., Zhao, H., Lu, X., Hu, X., & Ma, Z. (2018). A novel feature extraction methodology for sentiment analysis of product reviews. *Neural Computing and Applications*, 1-18.

[16]. Blunsom, P., Kocik, K., & Curran, J. R. (2006). Question classification with log-linear models. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 06. doi:10.1145/1148170.1148282

[17]. Ray, S. K., Singh, S., & Joshi, B. (2010). A semantic approach for question classification using WordNet and Wikipedia. Pattern Recognition Letters, *31*(13), 1935-1943. doi:10.1016/j.patrec.2010.06.012
[18]. Huang, Z., Thint, M., & Qin, Z. (2008). Question classification using head words and their hypernyms. Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP 08. doi:10.3115/1613715.1613835
[19]. Mollaei, A., Rahati-Quchani, S., & Estaji, A. (2012). Question classification in Persian language based on conditional random fields. 2012 2nd *International E-conference on Computer and Knowledge Engineering* (ICCKE). doi:10.1109/iccke.2012.6395395
[20]. Sherkat, E., & Farhoodi, M. (2014). A hybrid approach for question classification in Persian automatic question answering systems. 2014 4th *International Conference on Computer and Knowledge Engineering (*ICCKE). doi:10.1109/iccke.2014.6993377
[21]. Loni, B., Khoshnevis, S. H., & Wiggers, P. (2011). Latent semantic analysis for question classification with neural networks. 2011 IEEE Workshop on Automatic Speech Recognition & Understanding. doi:10.1109/asru.2011.6163971
[22]. Fauzi, M. A., Arifin, A. Z., & Yuniarti, A. (2017). Arabic Book Retrieval using Class and Book Index Based Term Weighting. *International Journal of Electrical and Computer Engineering* (*IJE*CE), *7*(6), 3705. doi:10.11591/ijece.v7i6.pp3705-3710
[23]. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. Weinberge (Ed.), *Advances in Neural Information Processing Systems*, *26*, 3111-3119.
[24]. Hanchuan Peng, Fuhui Long, & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226-1238. doi:10.1109/tpami.2005.159
[25]. Kennedy, J., & Eberhart, R. (n.d.). Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks. doi:10.1109/icnn.1995.488968
[26]. Riget, J., & Vesterstrøm, S. J. (2002). A Diversity-Guided Particle Swarm Optimizer– the ARPSO. EVALife Technical Report no. 2002-02, Aarhus Universitet, Department of Computer Science, Aarhus
[27]. Krink, T. A., Vesterstrom, J., & Riget, J. (2002). Particle swarm optimisation with spatial particle extension. Proceedings of the Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600). doi:10.1109/cec.2002.1004460
[28]. Hendtlass, T., & Randall, M. (2001). A Survey of Ant Colony and Particle Swarm Meta-Heuristics and Their Application to Discrete Optimization Problems. In Proceedings of the Inaugural Workshop on Artificial Life, 15-25.
[29]. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, *1*(1), 81-106. doi:10.1007/bf00116251.

How to cite this article: Kaur, R. and Bhardwaj, V. (2019). Gurmukhi Text Emotion Classification System Using TF-IDF and N-gram FeatureSet Reduced Using APSO. *International Journal on Emerging Technologies*, **10**(3): 352–362.