Emotion Analysis of Twitter Data using Hybrid Support Vector Machines and Ant Colony Optimization Variant

[1]Sandeep Kaur, [2]Dr. Vinay Chopra
[1]Research Scholar, [2]Assistant Professor
[1]DAVIET,Jalandhar,
[2]DAVIET Jalandhar

Abstract : **With the widespread usage of social media, forums and blogs, customer reviews and opinions emerged as critical factor for decision making in various areas. A lot of researchers focused on automatically categorizing reviews on the basis of polarity such as positive, negative and neutral called as sentiment analysis. In this paper, framework for analysing emotions of users in social media text such as twitter using emotion theories has been presented. However, data collected from twitter contains lot of irrelevant and redundant features. Hence, it is important to select only those features which contribute best to the accuracy of the classifier. To tackle with such problem, a hybrid algorithm using Ant Colony Optimization variant Max-Min Ant system and Support Vector Machines has been introduced through this paper. It has been found that accuracy of classifier enhances when optimal features are selected using feature selection algorithm.**

IndexTerms **– Support Vector Machines, Feature Selection, Ant Colony Optimization, Max-Min Ant System.**

I. INTRODUCTION

Emotions play a vital role in everyday human-human interaction and human computer interaction as well. Emotion is a key factor that influences human behaviour which includes reasoning, decision making and interaction. Emotion analysis is a subfield of sentiment analysis which aims at making decisions from the emotions present in social media, blogs, online forums in the form of text. The basic task of emotion analysis is to extract emotions such as joy, anger, surprise, disgust, sadness, love and so on [2]. However, sentiment analysis focuses on classifying text on the basis of polarity such as positive or negative. Due to rise of social media from past two decades, sentiment analysis and emotion analysis witnessed ample amount of interest from scientific community [1,3].

Feature Selection is a major research topic for the development of classification methods. FS (subset selection or attribute selection) is the automatic selection of attributes in dataset that are most relevant for application of a learning algorithm [4] . The best subset is selected that contains least number of dimensions contributing well to the accuracy of the learning classifier. Hence, rest of the features and irrelevant dimensions are discarded and only best ones are kept. FS is usually done in the pre-processing stage and is very efficient technique to deal with noisy and redundant features and tend to keep only those features which are best suited for the classifier. There are basically two methods by which FS can be carried out:

1. Forward Selection: In this type of selection, initially there are no variables and variables are added gradually at each step so as the error is decreased. This process halts when further addition does not decrease the error rate.
2. Backward Selection: This is contrasting to the prior one. In this case, the selection starts with considering all the variables and removing one by one at each step until any further removal increases the error considerably.

A typical feature selection process consists of four basic steps as shown in Figure 1. The first step is subset generation which is a search procedure that produces candidate feature subsets for evaluation. After that, each candidate subset is evaluated and compared with the previous best subset according to a certain evaluation criterion. If the new subset is found to be better than the previous one, previous best subset is replaced. The process of subset generation and evaluation continues repeating until a given stopping criterion is satisfied. Finally, result is validated by prior knowledge. Feature selection has successfully been applied in many fields such as classification, clustering, association rules and regressions. FS algorithms broadly fall into three main categories: the filter approach, the wrapper approach and the hybrid approach. In the filter approach, feature subsets are selected and evaluated without requiring any classifier. In case of wrapper approach, one predetermined classifier's performance is used as evaluation criterion. In this approach, features are selected in such a way that subset improves the performance of the classifier. However, wrapper approach is proved to be more computationally expensive than the filter approach. On the other hand, hybrid approach considers taking advantage of the two approaches by exploiting their different evaluation criteria in different search stages.

In this paper, we have introduced Hybrid Max-Min Ant System and Support Vector Machines for Emotion analysis. Additionally, we have used lexicons as feature groups.

II. RELATED WORK

Emotion Analysis is concerned with the identification and classification of emotions present in the text. Social media is the platform which is used to express feeling, views or opinions related to any entity. Much research work has been carried out in the field of emotion analysis and sentiment analysis. Dandan Jiang et al. (2017) [5] proposed an innovative method Word Emotion

Association Network (WEAN) to do emotion extraction and sentiment computing of news event. The proposed method consists of two parts: Word Emotion computation through Word Emotion Association Network and word emotion refinement. In the word emotion computation phase, microblogs with emoticons are considered to calculate the corresponding emotion present in the microblog. For refinement of the emotions derived from the first phase, they used standard sentiment thesaurus. For testing, they used Malaysia Airlines MH370 news event as dataset and computed six basic types of emotions: love, joy, anger, sad, fear and surprise. Dario Stojanovski et al. (2015) [6] exploit a convolutional neural network architecture for emotion analysis in Twitter messages related to sporting events on 2014 FIFA world Cup. In this paper, seven different kinds of emotions were evaluated using hashtag labeled tweets that were collected from Twitter Streaming API. The training of the network is performed on two samples containing 1000 and 10000 tweets on which this approach achieves 50.12% and 55.77% accuracy respectively. Moreover, they have presented the analysis of this approach on three different games that have great impact on Twitter users. Cagatay Catal et al. (2017) [7] exploit a sentiment classification model based on Vote ensemble classifier utilizes from three individual classifiers: Bagging, Naïve Bayes and Support Vector Machines. Moreover, in bagging they used SVM as base classifier. The main focus of this research is to improve the performance of machine learning classifiers for sentiment classification of Turkish reviews and documents. Their experimental results show that multiple classifier system based approaches are much better for sentiment classification of Turkish documents. They performed experiments on three different domains such as book review, movie reviews and shopping reviews. The authors concluded that this approach is not restricted to just one domain and can be extended to several other domains as well. Nadia Abd-Alsabour et al. (2010) [4] presented an ant colony optimization approach for feature selection for pattern classification. The proposed ACO differs from most conventional ACO because they used heuristic information in order to guide the search process besides the pheromone values. The efficiency of the proposed algorithm was tested on ten artificial and real world datasets. Moreover, they tested the effect of feature selection algorithm on the basis of classifier. Two different experiments were conducted. In first experiment, all features were considered and SVM was used for pattern classification. In the second one, the proposed FS-SVM algorithm was used that calculated best feature subset. By and large, they achieved promising results in terms of solution quality and number of selected features.

III. Feature Selection Using Max-Min Ant System(MMAS)

In this section Ant Colony Optimization variant Max-Min Ant System is explained. The difference between Ant Colony system and Max-Min Ant System is that in case of MMAS, global best ant or iteration best ant is used to update pheromone.

Initialization Initially, population of ants and intensity of pheromone trail associated with any feature is determined. Moreover, maximum number of allowed iterations is defined.

Heuristic Desirability In ACO algorithm [8], constructive heuristic is a basic requirement for probabilistically constructing solutions. A solution construction is empty in the beginning and solutions are assembled as sequences of elements from finite set of solution components. After that, a feasible solution component is added to the current partial solution at each construction step. Heuristic desirability of choosing between features could be any subset evaluation function. In the proposed algorithm, CFS subset evaluation is used as heuristic desirability. Then the probability that ant k will include feature i in its solution is called as probabilistic transition rule and is given by the following equation:

$$P_{ij}^{k}(t) = \begin{cases} \dfrac{[\tau_{ij}(t)]^{\alpha}.[\eta_{ij}]^{\beta}}{\sum_{l\epsilon N_i^k}[\tau_{il}(t)]^{\alpha}.[\eta_{il}]^{\beta}} & if\ j \epsilon N_i^k \\ 0 & otherwise \end{cases} \quad (3.1)$$

where η_{ij} is a heuristic value, $\tau_{ij}(t)$ is pheromone trail value, α and β are two parameters which determine the relative influence of the pheromone trail and the heuristic information, and N_i^k is the feasible neighborhood of ant k when being at node i.

Update Pheromone Unlike, Ant Colony system, in this approach only one single ant which is best-so-far ant is allowed to deposit pheromone and update pheromone trails. As a consequence, the pheromone update rule is given by the following rule

$$\tau_{ij}(t+1) = (\rho).\tau_{ij}(t) + \Delta\tau_{ij}{}^{best} \quad (3.2)$$

where $\Delta\tau_{ij}{}^{best} = 1/f(s^{best})$ and $f(s^{best})$ denotes the solution cost of either the iteration-best (s^{ib}) or the global-best solution (s^{gb}) . ρ is the pheromone evaporation constant.

Pheromone trail limits If independent of the choice between iteration-best and global-best ant for updating the solution, search stagnation may occur due to the reason that pheromone trail is significantly higher for one choice than for other. Therefore, ants will prefer the same solution component over other alternatives and will construct the same solution over and over again. [9] Thus, the exploration of the search space stops. To avoid this search stagnation, it is important to influence the probabilities for choosing the next solution component which depend on the pheromone trails and the heuristic information. Max-Min Ant System imposes explicit limits τ_{min} and τ_{max} to avoid the relative differences between the pheromone trails on the minimum and maximum pheromone trails $\tau_{ij}(t)$ such as

$$\tau_{ij}(t) = \begin{cases} \tau_{max} & ,if\ \tau_{ij}(t) > \tau_{max} \\ \tau_{min} & ,if\ \tau_{ij}(t) < \tau_{min} \\ \tau_{ij} & otherwise \end{cases} \quad (3.3)$$

Solution Construction The overall process of feature selection begins by generating a number of ants which are then placed randomly on the graph i.e. each ant starts with one random feature. From these initial positions, they traverse nodes probabilistically until a traversal stopping criterion is satisfied. The resulting subsets are gathered and then evaluated. If an optimal subset has been

found or the algorithm has executed a certain number of times, then the process halts and outputs the best feature subset encountered. If none of these conditions hold, then the pheromone is updated, a new set of ants are created and the process iterates once more (Figure 3.1)[8-9].

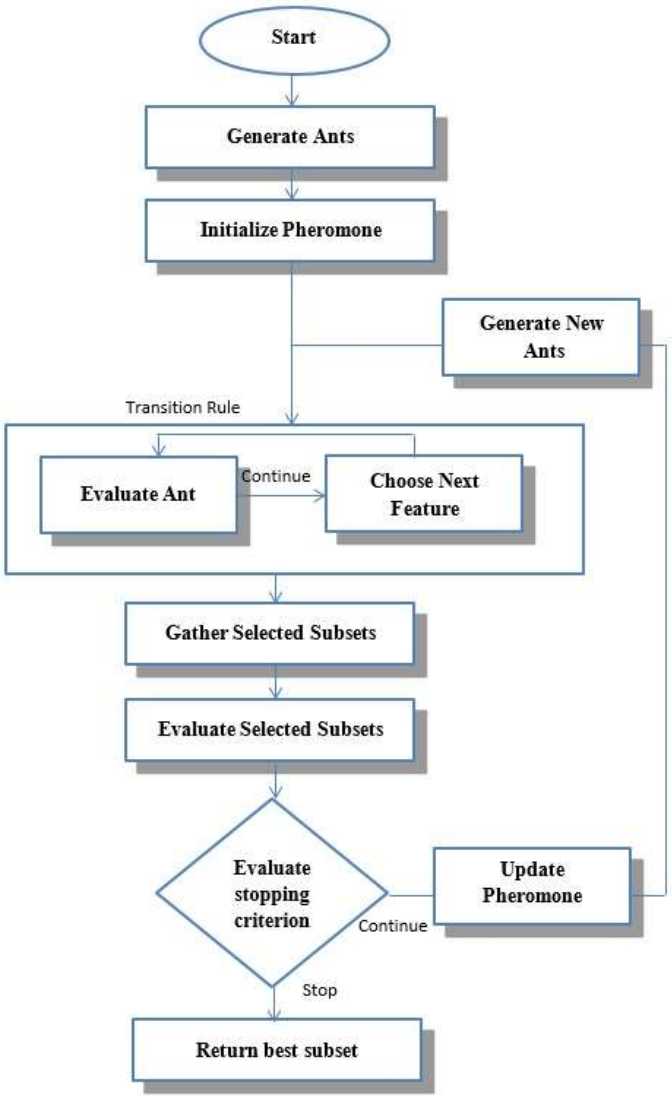

Fig 3.1 Flowchart of Ant Colony Feature Selection

IV. PROPOSED METHODOLOGY

4.1 Data Acquisition and Annotation: The first and foremost step is to collect data for emotion analysis. Ekman's six basic emotions are considered to represent affect and one more emotion (mixed emotion) is added.

i. Training dataset for four emotions labeled as Fear, Anger, Joy and Sadness are collected from WASSA-2017 Shared Task on Emotion Intensity (https://arxiv.org/abs/1708.03700).

ii. Secondly, tweets for emotions labeled as disgust, surprise and mixed emotion have been collected from twitter using Twitter API by searching hashtags such as #sickness #disgusting #incredible and so on. Annotation has been done manually by taking seed words from WordNet Affect.

4.2 Preprocessing and Filtration: Preprocessing and Filtration is performed to normalize the data. Filtering techniques like StringToWordVector filter, which converts String attributes into a set of attributes representing word occurrence information from the text contained in the strings, has been used. Some other pre-processing steps applied are Tokenization, Stop Words removal, dimensionality reduction etc.

4.3 Feature Groups: Feature sets are defined for automatic classification of emotions in tweets, it is important to consider emotional words which distinctly characterize emotions in tweets with hashtags.

i. **NRC Hashtag Emotion Lexicon:** which provides association of words with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) generated automatically from tweets with emotion-word hashtags such as #happy and #anger.

ii. **BingLui:** counts the number of positive and negative words from the Bing Liu lexicon.

iii. **AFINN:** calculates positive and negative variables by aggregating the positive and negative word scores provided by this lexicon.

4.4 Classification: Classification has been performed using hybrid Max-Min Support Vector Machines.

4.5 Performance Evaluation: Evaluation of performance has to be done using metrics like accuracy, precision, recall, F-measure, TP Rate, FP Rate and then the results of Support Vector Machines and Hybrid Max-Min Ant System – Support Vector Machines are compared on the basis of these parameters.

IV. RESULTS AND DISCUSSION

For comparison of performance of classification using SVM and MM-SVM, three parameters accuracy, precision and recall have been calculated and Table 1 presents the values of these parameters for both SVM and MM-SVM. In this work, unigrams are used for feature extraction and Term frequency inverse document frequency for feature weighting have been used. Moreover, feature vectors have been created using feature group. Ads ascertained from table 1, average accuracy, precision and recall obtained with MM-SVM system achieved much better than using SVM for the same dataset. The comparison is presented in Figure 2.

Table 1. Performance analysis of SVM and Hybrid MM-SVM

Parameter	SVM	MM-SVM
Accuracy	71.17%	75.31%
Precision	0.715	0.761
Recall	0.712	0.753

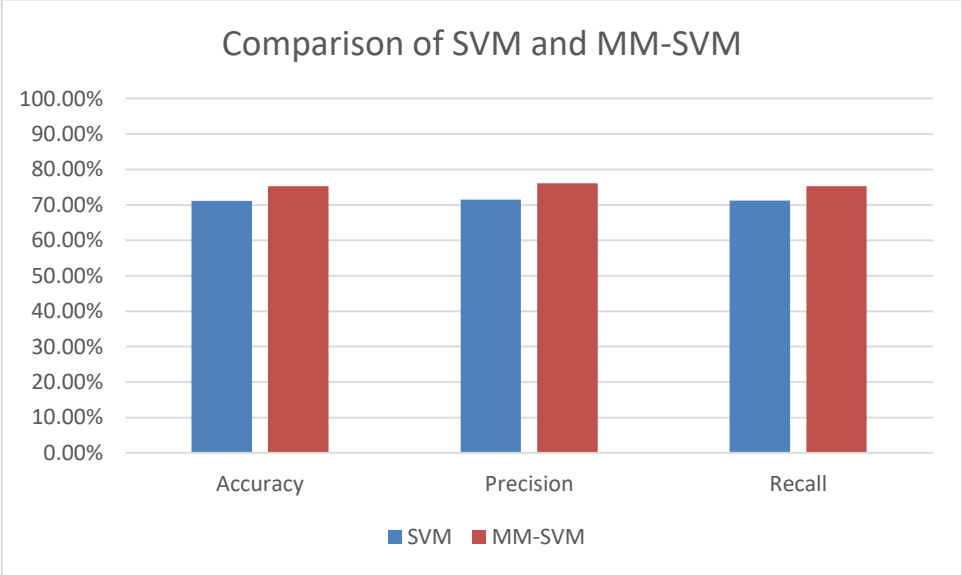

Fig 2. Comparison of SVM and Hybrid MM-SVM

V. CONCLUSION

It is observed that emotion analysis has become an important and essential area of text mining and has attracted lots of research interest over the past decade. Though there are many algorithms and techniques that are available to analyse the emotions present in the text, there are less techniques that deal with feature selection in analysing emotions.

In this study, we proposed a hybrid algorithm using Max Min Ant System and Support Vector Machines for feature selection and classification in emotion analysis while at the same time improving the accuracy of extracting emotions using twitter dataset. Our proposed framework for emotion analysis is compared with Support Vector Machines without feature selection algorithm on twitter dataset and results have shown that classification by using this efficient method has improved the accuracy.

More future research is needed to solve the emotion analysis challenges in the form of intensity of emotions present in the text in the range of 0 and 1. Moreover, sarcasm can be detected present in the text.

REFERENCES

[1] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79-86. Association for Computational Linguistics, 2002.

[2] Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169-200.

[3] D'Mello, Sidney, Rosalind W. Picard, and Arthur Graesser. "Toward an affect-sensitive AutoTutor." *IEEE Intelligent Systems* 22, no. 4 (2007).

[4] Abd-Alsabour, N., & Randall, M. (2010, December). Feature selection for classification using an ant colony system. In e-Science Workshops, 2010 Sixth IEEE International Conference on (pp. 86-91). IEEE.

[5] Jiang, Dandan, et al. "Sentiment Computing for the News Event Based on the Social Media Big Data." *IEEE Access* 5 (2017): 2373-2382
[6] Stojanovski, Dario, et al. "Emotion identification in FIFA world cup tweets using convolutional neural network." *Innovations in Information Technology (IIT), 2015 11th International Conference on*. IEEE, 2015.
[7] Catal, Cagatay, and Mehmet Nangir. "A sentiment classification model based on multiple classifiers." *Applied Soft Computing* 50 (2017): 135-141.
[8] Dorigo, M., Maniezzo, V., & Colorni, A. (1996). Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *26*(1), 29-41.
[9] Stutzle, T., & Hoos, H. (1997, April). MAX-MIN ant system and local search for the traveling salesman problem. In *Evolutionary Computation, 1997., IEEE International Conference on* (pp. 309-314). IEEE.