MULTI ANT COLONY OPTIMIZATION FOR OPINION CLASSIFICATION

Dr.K.Saraswathi, Dr.N.T.Renukadevi, Dr.S.Karunakaran

Abstract: Opinion Mining could be a sort of process in Natural Language Processing to track the disposition or supposition of individuals about a particular item, subject or service. This can be too called as Opinion Investigation or sentiment analysis which includes building a framework to gather and look at the suppositions, feelings, around the item, subject and administrations made in web journal posts, comments, surveys or tweets. This paper thinks about opinion and opinion based classification for movie surveys. To begin with, feature extraction has been done with Inverse document frequency strategy. Information Gain based feature selection process has been done from the reviews for effective feature selection. At long last, Multi Objective function based Ant Colony Optimization (MOFACO) procedure has been utilized for viable classification of surveys with optimized feature selection strategy. This inquire about work accomplishes the classification accuracy of 90.89%.

Keywords: Support Vector Machine, Naive Bayes, Ant Colony Optimization, Inverse Document Frequency, Information Gain.

1. INTRODUCTION

Opinion mining is a field of study that analyzes people’s sentiments, evaluations, appraisals, opinions, attitudes and emotions towards products, services, individuals, organizations, issues, events, topics and their attributes. It represents a big problem space. It also has many names and slightly different tasks, e.g., sentiment analysis, opinion extraction, sentiment mining, affect analysis, subjectivity analysis, emotion analysis, review mining and so on. Opinion mining helps to collect opinions about the positive and negative aspects of a particular topic. The opinions collected are analyzed then the positive and highly scored opinions obtained about a particular product are recommended to the users to make decisions on products. Large companies and business people are also making use of opinion mining to improve marketing. Furthermore, politicians keep on changing their campaign policies based on the people’s expectations[1]. The process in opinion mining is specified in two ways as Input text pre-processing or transformation of opinionated texts into lexical units which can be processed, and opinion classification which represents the calculation of the polarity (positive or negative) of a subjective opinion. The overall sentiment of a document is not necessarily the sum of the content parts. This phenomenon is one main reason why machine learning approaches fail under some situations[1]. Sentiment Classification can be done by using Supervised Learning methods and unsupervised learning methods. In supervised learning, each example is a pair consisting of an input object and the desired output value. Some classification algorithms using supervised learning are: Logistic Regression, Decision trees, Support Vector Machine (SVM), k-Nearest Neighbors, Naive Bayes, Random forest, Linear regression and polynomial regression[2].

In Sentiment Classification using Unsupervised Learning commonly in data mining, the problem of an unsupervised learning task is trying to find hidden structure from unlabeled data. Since the examples given to the learner are unlabeled, there is no sample or model to evaluate a potential solution. Some of unsupervised learning algorithms in classification of data in data mining are: Clustering, Neural Networks and anomaly detection. Unsupervised learning is based originally on sentiment words. It performs classification based on some fixed syntactic patterns that are likely to be used to express opinions. These syntactic patterns are composed based on POS tags. POS defines that for a given sentence, the part of speech for each word is determined. Many words, especially common ones, can serve as multiple POS. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb or adjective and "out" can be any of at least five different parts such as noun, verb, adverb, adjective and pronoun of speech. Words of different POS may be considered differently. In opinion mining, adjectives are important indicators of opinions. The Penn Treebank POS tags are mostly use Standard POS tags. POS tags compute semantic orientation of documents based on aggregated semantic orientation values of the selected opinionated POS tags extracted from the review document. A number of unsupervised learning approaches take the credit of creating a sentiment lexicon first in an unsupervised manner, and then determining the degree of positivity (or subjectivity) of a text unit via some functions based on the positive and negative (or simply subjective) indicators, as determined by the lexicon, within it[3]. Feature selection in opinion mining is an important research field and it develops machine learning and it has been utilized in data mining for years and is now applied to fields like text mining, genomic analysis, intrusion detection and image retrieval. When new applications emerged, many challenges also arose needing new theories and methods to address high dimensional/complex data. Optimal redundancy removal, stable feature selection and auxiliary data and prior knowledge exploitation in feature selection are the primary and challenging problems in feature selection. Feature selection and feature reduction attempt to decrease the dimensionality, i.e., the number of features, for the classification task. The classification phase of the process finds the actual mapping between patterns and labels or targets. Active learning, a kind of machine learning, is a promising way for sentiment classification to reduce the annotation cost. So far, a large volume of literature has been published on the direction of feature selection for the classification. Optimization for feature selection would lead to the genetic approach which has a little chance to get stuck at local minima. Genetic approaches are finding widespread applications in solving problems requiring efficient and effective search in the synthesis of neural network architectures, scheduling, numerical optimization and so on and results in solutions that are globally optimal or nearly to the optimal. Machine
learning feature selection is a global optimization problem that reduces feature count, eliminates irrelevant, noisy and redundant data and resulting in recognition accuracy.

The rest of the paper is organized as follows: Section 2 includes Literature Review, Section 3 discuss on feature extraction techniques, Section 4 deals with feature selection and Section 5 includes the experimental results and Section 6 is the conclusion.

2. LITERATURE REVIEW

A medical opinion lexicon for mining health reviews available on different health forums. This technique works based on the incremental modal and corpus of health reviews by creating medical polarity lexicon for medical terms. In each increment, the vocabulary of lexicon is enhanced systematically, polarity score with each word is attached, and finally, the resulting lexicon is filtered from unnecessary words by using word sense disambiguation techniques. The comparative results show the efficiency of the developed method with an accuracy of 82% on training corpus and 78% on testing corpus of health reviews[4]. Various techniques have been developed for the key tasks of opinion mining. Then an overall picture of what is involved in developing a software system for opinion mining on the basis of the survey and analysis is presented with several existing tools and methods in opinion mining[5]. It has been described techniques and approaches to directly enable opinion-oriented information seeking systems. An attempt has been made to deal with various approaches to be attempted on a computational treatment of sentiments and opinions. Various supervised or data-driven techniques for opinion mining like Naïve Bayes, Maximum Entropy (ME), and SVM are elaborated and their strengths and drawbacks are touched upon. From the survey, the final result obtained is unigram feature selection with SVM produces better results comparing with Naïve Bayes and maximum entropy[6]. Similarly, the proposed Sentiment Probabilistic Latent Semantic Analysis (S-PLSA), in which a review is considered as a document generated by a number of hidden sentiment factors in order to capture the complex nature of sentiments. Training an S-PLSA model enables to obtain a succinct summary of the sentiment information embedded in the reviews. Based on S-PLSFA, the study has proposed an Auto-Regressive Sentiment-Aware (ARSA) model for sales prediction. Further the accuracy of prediction is improved by considering the quality factor, with a focus on predicting the quality of a review in the absence of user-supplied indicators, and presented an Auto-Regressive Sentiment and Quality Aware model (ARSOA), to be utilized with sentiments and quality for predicting product sales performance[7]. An Artificial Immune System (AIS) technique is used to identify Malaysian online movie reviews. This opinion mining process uses three string similarity functions, namely cosine similarity, Jaccard coefficient and Sorensen coefficient. In addition, AIS performance is compared with the other traditional machine learning techniques like SVM, Naïve Bayes and k-Nearest Network. The results of the findings are elaborated which shows that AIS technique with k-NN produces improved accuracy up to 15% than others [8]. The state-of-the-art supervised machine learning methods for sentiment analysis are evaluated. Different pre-processing techniques are explored and various features and classifiers are employed. Five different feature selection algorithms and the influence of named entity recognition and preprocessing on sentiment classification performance is investigated. In addition to newly created social media dataset, the results for other popular domains, such as movie and product reviews are reported. This would not only extend the current sentiment analysis research to another family of languages but also encourage competition, leading potentially to the production of high-end commercial solutions[9]. Similarly, Consento method was used as a consensus search engine developed to answer subjective queries. Consento performs segment indexing, as opposed to document indexing, to capture semantics from user opinions more precisely. In particular, the study has defined a new indexing unit named as Maximal Coherent Semantic Unit (MCSU). An MCSU represents a segment of a document and it captures a single coherent semantic. A new ranking method, called Consensus Rank that counts online comments referring to an entity as a weighted vote is also introduced. To validate the efficacy of the proposed framework, Consento with standard feature retrieval models and their recent extensions for opinion based entity ranking is compared. Experiments with movie and hotel data are carried out which show the effectiveness of the framework[10]. An architecture that can be utilized to automatically analyze the sentiments of the messages. This system is combined with manually annotated data from Twitter, one of the most popular micro blogging platforms, for sentiment analysis. In this system, machines can learn automatically extracting the set of messages which contain opinions, filter out non-opinion messages and determine their sentiment directions, i.e., positive or negative. Experimental results confirm the effectiveness of the system on sentiment analysis in real micro blogging applications[11]. A review rating prediction method was introduced by incorporating the character of reviewer's social relations as regularization constraints, into content based methods. In addition, a new classification method is advanced to classify the social relations of reviewers into strong social relation and ordinary social relation. For strong social relation of reviewers, higher weight is given than ordinary social relation when incorporating the two social relations into content based methods. Experiments on two real movie review datasets demonstrate that the method of taking different social relations has better performance than the method of considering social relations as a whole and the content based methods[12]. The features are extracted from set of reviews using TF-IDF and the reviews are classified into positive or negative using bagging algorithms. This work is evaluated using a subset of IMDB. Bagging algorithm produces better classification accuracy and reduced Root Mean Squared Error (RMSE)[13]. The current trends of market in terms of relevant product features are taken. The users’ interest towards these features is extracted by mining their opinions. Subsequently, market segmentation is done by clustering similar users and the best segment(s) is selected for product promotion. This strategy finally exploits the social connectedness among online users to identify the best initial seeds. This approach is also capable of attracting the attention of a large span of web users by employing a small fraction of advertising budget and it has
the potential in current e-marketing scenario[14]. A concept extraction algorithm based on a novel concept parser scheme is used to extract semantic features that exploit semantic relationships between words in natural language text. Additional conceptual information of a concept is obtained using the Concept Net ontology. Concepts extracted from text are sent as queries to Concept Net to extract their semantics. Important concepts are selected while redundant concepts are eliminated using the MRMR feature selection technique. All the selected concepts are then used to build a machine learning model to classify a given document as positive or negative. The study evaluates the concept extraction approach using a benchmark movie review dataset provided by Cornell University and product review datasets on books and electronics. Comparative experimental results show that the proposed approach to sentiment analysis performs better than the existing state-of-the-art methods[15].

3. PERFORMANCE EVALUATION OF CLASSIFIERS USING INFORMATION GAIN
Opinion mining is a promising discipline, which is defined as the combination of information retrieval and computational linguistic techniques. It deals with the opinions expressed in a document. Its main goal is solving the problems related to opinions about products, politics in newsgroup posts, review sites and so on. There are several techniques available for summarizing and categorizing customer reviews like data mining, Information Retrieval, Text Classification and Text Summarization[16]. Customers can post reviews on web communities, discussion forums, twitters, blogs and products’ websites. These comments or reviews are called user generated contents. Several tools play a vital role in extracting data source in opinion mining. It facilitates users to know and to retrieve information about the product from the other customers’ reviews, which have been used instead of asking the opinion with friends and families. Companies can extract opinionated text from product websites instead of conducting surveys and hiring the external consultants to know about the clients’ opinions. The final step of any classification is to carry out the evaluation of the proposed algorithm using standard techniques. Likewise, the main motive of evaluation in any machine learning based classification algorithm is to determine the usefulness of the learning classifier on different data sets. The properties of any learning method used such as accuracy, precision and recall to measure the process of evaluation[17].

3.1. Feature Extraction
Features are extracted using Term Frequency-Inverse Document Frequency (TF-IDF) for document classification. A list of stop words (commonly occurring words) and stemming words (words with similar context) is also prepared. The term frequency (tf), which includes a number of documents having the term, is computed. TF-IDF is a statistical method to index the terms based on their importance. In which, term vector is utilized to represent term presence and term frequency. The TF is defined as in Equation (1).

$$tf (t, d) = 0.5 + \frac{0.5 \times f (t, d)}{\text{maximum occurrences of words}}$$ (1)

Here ‘t’ represents the term and ‘d’ denotes documents. Rarely occurring words/terms are more informative than the terms which occur frequently. Hence rare words are assigned higher weights than those terms used regularly. These are captured by document frequency term t (df_t) and inverse document frequency (idf_t) represents scaling factor. Term t’s importance is scaled down when used frequently. The idf_t is defined as in Equation (2).

$$idf (t, d) = \log \left( \frac{p}{\text{no. of documents term t appears}} \right)$$

Equation (2)
Here ‘|D|’ represents total number of documents. Feature Extraction has been performed by using Term Frequency-Inverse Document Frequency Technique. The tf-idf weight is a weight that is used in information retrieval and text mining. This weight is a statistical measure used to estimate the importance of a word is to a document. The importance measured as high proportionally to the number of times a word occurs in the document but is offset by the frequency of the word in the corpus. The total TF-IDF weight for a token in a document is the product of its TF and IDF weights. It is specified in Equation (3).

$$tfidf (t, d , D) = tf(t, d) \times idf(t, d)$$ (3)

The TF-IDF method extracts the features from the reviews but it contains large number of feature subsets. Obviously, it leads to more operating time and ambiguity in the results. So, the feature selection methods are applied.

3.2. Feature Selection Using Information Gain
Feature selection is an activity which selects relevant and reduced features based on a particular measurement. Its main purposes are to simplify the process of training and to reduce the time of training process[18]. The performance of certain classifiers like k Nearest Neighbor and Bayesian algorithms produce poor when the features are too many. However, it is important to adopt a feature selection technique which reduces the number of features without reducing the performance of opinion mining. Some common feature selection techniques such as POS, IG, Document Frequency and Chi Square are incorporated in previous researches on opinion mining. There are three groups of feature selection techniques, i.e. filter, wrapper and embedded. In a filter category, a particular mathematical equation is used for selecting a group of features and it may be used with any classifier. In contrast, wrapper works as the features that are selected and the embedded techniques are bound to a particular classifier. Other than being very rigid in terms of classifier, the wrapper and embedded techniques normally require for high classification accuracy[19]. IG is a feature ranking method based on decision trees that produces reduced features and exhibits good classification performance. The goal of IG is to select features that reveal the most information about the classes. Each feature has been assigned an own IG value which determines whether this feature is to be
selected or not for the classification task. A threshold value is used for checking the features; if a feature has a greater IG value than the threshold value, then the feature is chosen; otherwise, it is not selected. Let $S$ be the set of $n$ instances and $C$ is the set of $k$ classes. $P(C_i, S)$ represents the division of the example in $S$ that has class $C_i$. Then the expected information from this class membership are derived as given by Equation (4).

$$Info(S) = -\sum_{i=1}^{k} P(C_i,S) \times \log(P(C_i,S))$$  \hspace{1cm} (4)

If a particular attribute $A$ has $v$ distinct values, the expected information is attained by the decision tree in which $A$ is the root, and the weighted sum of expected information of the subsets of $A$ is based on the distinct values $v$. Consider $S_i$ is the set of instances and $A_i$ is the value of attribute $A$ in Equation (5).

$$Info_A(S) = -\sum_{i=1}^{v} \frac{|S_i|}{|S|} \times Info(S_i)$$  \hspace{1cm} (5)

Then the difference between $Info(S_i)$ and $Info_A(S)$ afford the information gained by partitioning $S$ according to the test $A$ in Equation (6).

$$Gain(A) = Info(S) - Info_A(S)$$  \hspace{1cm} (6)

A higher information gain would result in a higher likelihood of obtaining pure classes in a target class.

4. MULTI OBJECTIVE FUNCTION BASED ANT COLONY OPTIMIZATION (MOFACO) TECHNIQUE

Opinion classification involves classifying the opinionated review text into two forms as positive or negative sentiment reviews. Machine learning is used for text classification with different classification algorithms such as Decision Tree, Naive Bayes, and SVM and so on. This approach is applied for review text classification of data such as movie, product, medical reviews and news reviews. Sentiment analysis is also done by the other methods such as dictionaries, word lexicons, word senses and so on.

Multi objective function is better because the feature selection in machine learning is a global optimization problem that reduces feature number, removes irrelevant, noisy and redundant data, resulting in recognition accuracy. It is an important step that affects the system performance of pattern recognition. Typically, feature selection problems are solved through the use of single objective optimization techniques like genetic algorithm[20]. This type of technique optimizes only a single quality measure; for example, recall, precision or F-measure at a time. Sometimes, it is unable to capture a good classifier’s quality reliably by a single measure. A good quality classifier must have recall, precision and F-measure values optimized simultaneously instead of doing parameter values alone. And also swarm intelligence technique is used widely to address the selection of optimal feature set.In customer review documents, reviewers convey positive, negative or both the sentiments about the objects and attributes. Document level and sentence level classification would not express the likes and dislikes of consumers about particular attributes of object. For example, if users comment on a mobile phone, they basically will comment upon Camera result, LCD size, speaker, weight and so on. For example, on camera output, 125 comments express the positive opinions and 25 comments may be negative. If a new customer is interested in camera quality of mobile, he or she can take decisions easily as to purchase the mobile or not. To explore the detailed opinion on a product or any topic or service, a detailed opinion mining study is required, which is called feature based opinion mining.

4.1 Ant Colony Optimization (ACO) for Feature Selection

The problem in feature selection lies in locating an optimal subset which is used to reduce computational overheads and to improve classification accuracy. A feature selection problem may be represented as an ACO-suitable problem by using the node for representing a feature and edge representing the cost function linking to the next feature in a graph. Ant colony algorithms are optimization techniques inspired by the foraging behavior of real ants in nature. When searching for food, ants primarily explore the area surrounding their nest in a arbitrary manner[21]. When an ant finds a food source, it estimates the quantity and quality of the food and brings the food back to the nest. During return, the ant deposits a chemical pheromone trail on the ground. It is easy to understand that the quantity of pheromone deposited, would depend on the quantity and quality of the food, and this pheromone deposited, will guide other ants to the food source. The indirect contact between the ants through pheromone trails enables them to find the shortest paths between the nest and food sources. This special characteristic of real ant colonies is exploited in artificial ant colonies in order to solve difficult combinatorial optimization problems. In Figure 1, the ant colony algorithm represents artificial ants probabilistically build solutions by taking into account dynamical artificial pheromone trails that is given in the steps A,B,C and D. The essential component of ACO algorithm is the pheromone model including the state transition rule and updating rule, which is used to probabilistically sample the search space.

![Fig. 1. ACO Performance](image)

ACO is suitable for multi objective optimization problems and it has been applied for several applications. The combination of the classifier parameters, $C$ and $\gamma$, based on RBF kernel of the SVM classifier is represented by using ant’s solution. The classification accuracy of the SVM classifier is utilized to direct the updation of solution archives. Based on the solution archive, the transition probability is calculated to choose a solution path for an ant.
The overall process to hybridize ACO and SVM is proposed[22].

1: Initialize pheromone trail
2: while stopping criteria not met do
3: for all ants do
4: Deposit ant randomly
5: while solution incomplete do
6: Select next element randomly according to the pheromone trail
7: end while
8: end for
9: Update pheromone trail
10: end while

In the current research work, the number of ants selected for ACO is equivalent to 15. Using ACO, the parameters of kernel functions are tuned. Termination condition is set to maximize the overall classification accuracy. SVM is implemented like SVM with poly kernel, SVM with RBF kernel (10,0.01), SVM with RBF kernel(100,0.1) and SVM with RBF kernel with ACO. MOFACO is utilized for better feature selection before applying the classification algorithms such as CART, Naïve Bayes and SVM. Finally ACO is used for parameter optimization. The focus of ACO optimization algorithm is to generate reduced size of salient feature subsets. ACO feature selection uses hybrid search technique combining wrapper/filter approaches. It modifies standard pheromone update and heuristic information measurement rules based on these approaches. It emphasizes the selection of salient features, and attaining a reduced number. It chooses a reduced number’s salient features using a subset size determination scheme. Termination condition maximizes the overall classification accuracy. Exponential fitness function was selected for faster convergence. The basis for any ACO algorithm is a constructive heuristic for probabilistic solution creation. A constructive heuristic assembles solutions as elements sequences from a finite solution components set. The basic thing for any ACO algorithm is a constructive heuristic for probabilistic solution creation. The function of a constructive heuristic is assembling solutions as element sequences from a finite solution component set. A solution construction begins with an empty partial solution, and at each production step, a current partial solution is prolonged by adding a feasible solution component from the solution component set. The probabilistic transition is represented by Equation (7):

\[
P_{i}^{k}(t) = \begin{cases} 
\frac{[\tau_{i}(t)]^{\alpha}[\eta_{i}]^{\beta}}{\sum_{i \in J^{k}}[\tau_{i}(t)]^{\alpha}[\eta_{i}]^{\beta}} & \text{if } i \in J^{k}, \\
0 & \text{otherwise},
\end{cases} 
\]

(7)

where \(J^{k}\) represents a set of feasible features that are added to a partial solution; \(\tau_{i}\) and \(\eta_{i}\) are pheromone value and heuristic desirability related to feature i respectively. \(\alpha\) and \(\beta\) are two arguments determining pheromone value’s and heuristic information’s relative importance. Pheromone evaporation on nodes is triggered after ants reach their solutions, and according to Equation (8), each ant \(k\) deposits a quantity of pheromone, \(\Delta \tau_{i}^{k}(t)\), on a node that it has used,

\[
\Delta \tau_{i}^{k}(t) = \begin{cases} 
\phi \tau(S^{k}(t)) + \frac{\varphi(n - |S^{k}(t)|)}{n} & \text{if } i \in S^{k}(t), \\
0 & \text{otherwise},
\end{cases} 
\]

where \(S^{k}(t)\) is a feature subset derived by ant \(k\) at iteration \(t\), and \(|S^{k}(t)|\) is its length. It is assumed that the performance of the classifier is equally important compared to subset length, and thus the values \(\phi = 0.5\), \(\varphi = 0.5\) are chosen. Pheromone is added using

\[
\tau_{i}(t+1) = (1 - \rho) \tau_{i}(t) + \sum_{k=1}^{m} \Delta \tau_{i}(t) + \Delta \tau_{i}^{k}(t) 
\]

(9)

where \(m\) is the number of ants at each iteration and \(\rho = (0.1)\) is a pheromone trail decay coefficient. In this, \(g\) indicates the best ant at each iteration[23].

5. EXPERIMENTATION AND RESULTS
5.1. IMDB Dataset
Experiments are carried out using the IMDB movie dataset and medical opinions collected from various websites. The movie dataset obtained in the public domain consists of experts’ opinions which is treated as the class label. For the medical blogs collected, three experts have rated their opinions and the best answer is selected by voting. The features are selected using IG. The top ranked features are selected and classified using MOFACO-CART, MOFACO-Naïve Bayes, MOFACO-SVM-Poly, MOFACO-SVM RBF (10,0.01), MOFACO-SVM RBF (100,0.1) and MOFACO-SVM-ACO methods are used. Table I and Figures 2 to 4 show the classification accuracy, recall and precision of IMDB dataset and medical blog dataset respectively.
Table I. Results for IMDB Dataset using multi objective function ACO

<table>
<thead>
<tr>
<th>Technique</th>
<th>Classification Accuracy</th>
<th>Recall for positive</th>
<th>Recall for negative</th>
<th>Precision for positive</th>
<th>Precision for negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOFACO-CART</td>
<td>80.89</td>
<td>0.8267</td>
<td>0.7911</td>
<td>0.7983</td>
<td>0.8203</td>
</tr>
<tr>
<td>MOFACO-Naive Bayes</td>
<td>81.89</td>
<td>0.8422</td>
<td>0.7956</td>
<td>0.8047</td>
<td>0.8345</td>
</tr>
<tr>
<td>MOFACO-SVM-Poly</td>
<td>82.33</td>
<td>0.84</td>
<td>0.8067</td>
<td>0.8129</td>
<td>0.8345</td>
</tr>
<tr>
<td>MOFACO-SVM RBF (10,0.01)</td>
<td>83.89</td>
<td>0.8578</td>
<td>0.82</td>
<td>0.8266</td>
<td>0.8522</td>
</tr>
<tr>
<td>MOFACO-SVM RBF (100,0.1)</td>
<td>88.67</td>
<td>0.8711</td>
<td>0.9022</td>
<td>0.8991</td>
<td>0.875</td>
</tr>
<tr>
<td>MOFACO-SVM-ACO</td>
<td>90.89</td>
<td>0.8978</td>
<td>0.92</td>
<td>0.9182</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Figure 2 Classification Accuracy for IMDB Dataset using multi objective function ACO

Figure 2 reveals that the classification accuracy of MOFACO-SVM-ACO is better by 11.64% than MOFACO-CART, by 10.42% than MOFACO-Naive Bayes, by 9.88% than MOFACO SVM-Poly, by 8.01% than MOFACO-SVM RBF (10,0.01), by 2.47% than MOFACO-SVM RBF (100,0.1) for IMDB dataset.

Figure 4 Precision for IMDB Dataset using multi objective function ACO

It can be observed from Figure 4, that the precision of MOFACO-SVM-ACO performs better by 13.97% & 9.26% than MOFACO-CART, by 13.17% & 7.55% than MOFACO-Naive bayes, by 12.16% & 7.55% than MOFACO SVM-Poly, by 10.49% & 5.45% than MOFACO-SVM RBF (10,0.01) and by 2.1% & 2.81% than MOFACO-SVM RBF (100,0.1) the precision for positive and negative by using IMDB dataset.

6. CONCLUSION

Opinion mining and sentiment analysis extract and classify the people’s opinions automatically from the internet sources. Classification is the most commonly used data mining technique. It is employed to predict the possible outcome from the given dataset on the basis of the defined set of attributes and a given predictive attributes. The present study uses SVM to classify as positive or negative feature sets from reviews extracted through the use of Term Frequency-Inverse Document Frequency. SVM classifies features using Polykernel and RBF kernel. A subset of
IMDB with 300 instances (150 positive and 150 negative) and medical opinion dataset are utilized for evaluation. Results show that the classification accuracy of IG-SVM RBF (100,0.1) performs better by 17.26% than IG-CART, by 14.91% than IG-Naïve Bayes, by 11.45% than IG bagging, by 8.43% than IG-SVM-Poly and by 6.87% than IG-SVM RBF (10,0.01) by using IMDB dataset.

REFERENCES


1. Dr. K.Saraswathi received the M.Sc., and M.Phil Degrees in Computer Science from Bharathiar University, India in 2004 and 2006 respectively. She received Ph.D. degree in Computer Applications from Anna University, India in 2018. Since 2008, she is working as an Assistant Professor in the Department of Computer Technology, Kongu Engineering College. Her research interests include Data Mining and Opinion mining

2. Dr. N.T. Renukadevi received the M.C.A and M.Phil Degrees in Computer Science from Bharathiar University, India in 2003 and 2008 respectively. She received Ph.D. degree in Computer Applications from Anna University, India in 2014. Since 2008, she is working as an Assistant Professor in the Department of Computer Technology, Kongu Engineering College. Her research interests include Data Mining and Image Processing. She is also a reviewer in reputed journals.

3. Dr. S. Karunakaran was born on March 15, 1979 in Erode. He received B.Sc., Degree in Computer Science from Nagamalai Navarasam Arts and Science College, Bharathiar University, Coimbatore, TamilNadu, India in 1999; Master of Computer Applications degree from Kongu Engineering College, Bharathiar University, Coimbatore, India in 2002 and M. Phil., degree in Computer Science from Manonmaniam Sundaranar University, Tirunelveli, India in 2003; Ph.D in Computer Science from Anna University Chennai in 2012. He became member of IEEE in 2007. Presently, he is working as an Assistant Professor (SRG), in the School of Computer Technology and Applications, at Kongu Engineering College, Perundurai, Erode, Tamilnadu, India. His academic interests include ad-hoc networks, distributed computing and wireless communications.