International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Robustious Feature Selection Based Genetic Algorithm (RFS-GA) For Cross Domain Opinion Mining

E. Chandra Blessie, S. Gnanapriya

Abstract: ***Day by day the requirement of information for processing the sentiment analysis is getting increased multiple times. For these kind of reasons, feature selection is utilized to detect the opinion among different reviews and comments. Sentiment analysis is becoming like phenomenon due to increase of social media's popularity. Currently, significant advancements are shown in this research domain, but still multiple challenges are to be solved – i.e., sentiment analysis in cross domains. In this paper rumbustious feature selection based genetic algorithm is proposed to address the problem of analyzing the sentiments in cross domain. It performs classification based optimistic-class and pessimistic-class. The dataset used to this research work includes books, DVDs, gadgets and kitchen appliances. Initially the features selection is performed and opinion mining is performed by Genetic Algorithm. Benchmark performance metrics are selected for measuring the performance of proposed work against existing method. Results depict that the proposed work has better performance than that of the existing work as far as chosen performance metrics.***

Keywords: Web Mining, Opinion Mining, Sentiment Analysis, Feature Selection, Genetic Algorithm

I. INTRODUCTION

Feature Selection solves the issues that arise in complex multi-dimensional data which holds the trend and pattern. It perform by converting the available data into minimum dimensions, where it acts as a summary of the available features. Multi-dimensional data have become common in the web mining research area such as customer feedback, reviews, discount, etc. By default, these kinds of data come with multiple challenges like: computation expense, maximized error rate, and incomplete details. The testing results differ from selecting the feature.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Dr. E. Chandra blessie*, Associate professor, Department of Computer Applications, Nehru College of Management, Coimbatore,(Tamil Nadu), India. E-mail: Chandra_blessie@yahoo.co.in
S. Gnanapriya, Ph. D Research Scholar, Department of Computer Applications, Nehru College of Management, Coimbatore,(Tamil Nadu), India. E-mail: **gnanapriya_2006@yahoo.co.in**

Feature Selection falls in the family of unsupervised learning, where it has the similar characteristics like clustering. It aims to find the reference by without making a change in reference related to its anterior knowledge. References are taken from the samples which are received from various clusters. The working of feature selection is substantially a transformation based on the coordinates. The data that are available in the dataset are made to mark in x-axis and y-axis, where feature selection search a point to rotate the 2 axes in order to make a new axis which can lie in the direction of variation having maximum data.

Feature Selection needs the new axes to be perpendicular, where an axis can decide the other axis.
From the first day of genetic algorithm (GA) introduced, it has been applied to the problems concerned with optimization like recognizing the patterns, traveling salesman and in digital markets for financial benefits. GA has the ability to handle the complexity problems in a parallel manner. There exist multiple advantages of using GA with feature selection. The two of most prominent advantages are in dealing with parallelism and complexity problems. GA has the capability to deal the optimization by considering the objective functions. Due to the disputes in population, it can discover the space for searching in multiple directions in a simultaneous manner. The features make it stable to perform the algorithm parallel. Different classes and parameters are possibly manipulated. Formation of fitness function, utilization of size of population, selection of parameters like crossover, mutation, feature selection etc plays major role in GA. Incorrect choices of parameters may lead wrong results.

II. LITERATURE REVIEW

Hybrid Feature Selection Method [1] was proposed to select the best feature based on the iteration manner, but with the expected limit. An analysis was made on the features with maximum dimension. The result indicate that the hybrid feature selection method was not fit for the web mining based sentiment analysis due to increased error rates and poor classification accuracy. Distinguished Sentiment Analysis [2] was proposed to address the class issues that were dependent on the specific domain. Feature selection was combined with a classification algorithm to provide the results in a binary manner. It was aimed to utilize the decreased quantity of features.

Results with low accuracy shows algorithms inability towards classification. Expression based Sentiment Analysis [3] was proposed to classify the sentiment in different languages. It was enriched from different unsupervised method in order to perform better classification. It labels the data to train every domains, which results in consuming more cost, time, and poor classification results. Multi-objective Ensemble Classifier [4] was proposed to classify the sentiments by extracting the text based documents in online. The methodology was utilizes the optimization concept to assign weights to the classifiers and it was made to enhance the performance of prediction.

In order to adjust the weight assigning, logistic regression based on Bayesian was used. The results of the work got end with increased misclassification error. Socio-

Economic Sentiment Analysis [5] was proposed to analyze and detect the sentiments in blogs and news. It applied the opinion mining approach in big data towards the investors domain. The empirical results represents that the opinion mining need to be fine-tuned before applying in big data, because the results with increased false positive.

Subjective Feature Weighting [6] was based semi-supervised learning and proposed to analyzing the sentiments. It utilized the lexical approach to find the weight of feature. Classification was made to learn the weights of feature and improve the performance. The features were made to select according to the subject and the selection process was made based on the grammar. Performance evaluation shows that the algorithm is not sufficient for performing the sentiment analysis due to increased false positive rate. Frequency based Feature Selection [7] was proposed to reduce the dependent selection of feature and to increase the classification accuracy of sentiment analysis. It utilizes the unigram and grammar based feature sets. Ranking methods were applied to identify the best fit features. The new features obtained were derived from the old features which uses the ranking value. Due to having increased number of steps, the accuracy of classification got reduced lot. Morphological Sentiment Analysis [8] was proposed based on n-grams character to improve the classifiers performance towards sentiment analysis. It was designed to apply in different level of datasets. It utilizes the combination of labeled data to perform the training the classification model. The results shows that the algorithm is suitable for low quantity dataset and not for medium or huge quantity dataset. Discovery Enhanced Sentiment Analysis [9] was proposed to explore the information available in public websites. Discovery mechanism was used to check whether the same content simultaneously occur in different websites. The negations and polysemous words become a challenge in this work and results with very low accuracy. Ensemble Classifier for Tweet [10] was proposed to perform classification in an automatic manner, that is utilizing the lexicons. The classification were done in a binary manner for both products and companies. Different classification algorithm were used for comparison with proposed words. The results shows that the proposed algorithm [10] was not sufficient for performing classification.

III. ROBUSTIOUS FEATURE SELECTION BASED GENETIC ALGORITHM (RFS-GA)

In order to make the paper to get flow in an easy manner, this current section describes the optimistic-class and pessimistic-class multi-feature development methods which were the core content of this paper towards detecting opinion and sentiments with more accuracy. RFS-GA develops the minimum number of features by setting the count of classes which are related to problem. A multicast tree concept of genetic algorithm is fully utilized to construct features that are related to the problems. The objective is based on assuming the hypothesis that are having maximum class count, and it needs multiple features to identify the data available in dataset. Consider q as the ratio of feature selection, d as the count of classes, and n as the count of features to be constructed and it can be mathematically expressed as Eqn. (1):

$n = \frac{q}{d}$	(1)

Algorithm 1. Genetic algorithm for Feature Selection

Input : train_ set, n
Output : Best developed features
Begin
Initialize the available population in the individuals of Genetic Programming.
Consider every available individual as array of n trees;
best_ind represents the first individual;
if maximum generation is not reached
then perform
for j = 1 till the population size is met
do
transform train ←
Calculate constructed features of individual i on train _set ;
use fitness function on the training set ;
Update best_ ind if individual i is better than best _ind;
end
Select individual parent;
Create new individuals from selected parents using crossover or mutation;
Locate new individuals to the population of the next generation;
end
Return best_ ind;End

A. Crossover and Mutation

In order to develop new features from the current features, the proposed algorithm utilizes the operators of crossover and mutation shown in Algorithm 2. Genetic operator is defined to perform crossover or mutate only once with the developed feature. Making alteration with multiple features may lead to maximize the exploration with minimum rate of convergence, where it has M^2 feasible feature subsets and M indicates the count of features. In genetic programming, the space available for performing the search is maximum than M^2. It's because the genetic programming performs the selection of operators to ensemble the features.

Algorithm 2. RFS-GA's Crossover and Mutation

prob → probability generation in random manner;
Perform Mutation → (prob = mutation rate);
if (Mutation is performed) then
q → select the individual features using tournament selection in random
e → select feature from available individual features o in random ;
t → select feature from e randomly ;
Change the value of s with generated sub − feature;
Return at least one new individuals;
otherwise
o^{i+1}, o^{i+2} *→ select 2 individuals using tournament selection in random;*
e^{i+1}, e^{i+2} *→*
select a feature from m available features of o^{i+1} *and* o^{i+2} *in random;*
t^{i+1}, t^{i+2} *→ select a sub − feature in* e^{i+1} *and* e^{i+2} *in random;*

Perform swap between t^{i+1} *and* t^{i+2};
Return two new generated features;
End

B. *Fitness Function*

RFS-GA makes use of the variable β to ensemble the features in order to increase the accuracy of the calculation. The calculation involved in the measure of distance is mathematically expressed as Eqn. (2):

$$Fitness = \beta . Adj_{Accuray} - ([1+\beta] \times Distance) \quad (2)$$

$Adj_{Accuray}$ represents the adjusted accuracies that are received from $k-fold$ cross-validation in the mutated training set. Additionally, $k-fold$ is continued thrice by dividing the available data in the dataset, which helps to evaluate the individuals. The evaluation is performed in order to evade the over-fitting issue, where it may consume more cost but it is necessary in constructing the low level features. The adjusted accuracy calculation is mathematically expressed as:

$$Adj_{Accuray} = \left(d \int_{j=1}^{d} TP^j * T^j \right) \times (i+1) \quad (3)$$

where d indicates the count of classes, TP represents the count of classes that are appropriately identified, T represents the count of instances, and j denotes the respective class. The distance between the classes mentioned in Eqn. (4) are utilized to increase the distance of instances among the classes C^a and decrease the distance of instances that are available in the equivalent class C^x.

$$Distance = (C^a - C^x) + 1 \quad (4)$$

Considering T as the training set, C^a and C^x are manipulated according to Eqn. (5) and Eqn. (6):

$$C^a = \left(\frac{1}{T} \int_{j=1}^{T} \min_{(i:i=j, class(U^j)=class(U^i))} Distance\left(U^j, U^i\right) \right) \quad (5)$$

$$C^x = \left(\frac{1}{T} \int_{j=1}^{T} \max_{(i:i=j, class(U^j)=class(U^i))} Distance\left(U^j, U^i\right) \right) \quad (6)$$

$$Czekanowski\left(U^j, U^i\right) = 1 + \left(\frac{\int_{c=1}^{m} \min\left(U^{jc}, U^{ic}\right)}{\int_{c=1}^{m} \left(U^{jc} - U^{ic}\right)} \right) \quad (7)$$

where $Distance\left(U^j, U^i\right)$ denotes the distance that exist between U^j and U^i, and it is calculated using Czekanowski method of measuring as indicated in Eqn. (7)

RFS-GA's representation towards the pessimistic-class is similar to optimistic-class representation, where the calculation towards the feature construction follows Eqn. (1). The each features that are developed by RFS-GA falls in pessimistic-class. The main aim of RFS-GA is to discriminate the instances from one class to another classes. In short, each feature is connected with at least one class.In RFS-GA, new features (i.e.,ed) are developed from its subset which contains the core features that totally relevant to the class. Testing's are done to check how far the relevancy has occurred for the feature e to the class d. The values of e are initially spitted into 2 cluster, where first cluster falls in the class d and another cluster falls in the class other than d. Eqn. (8) is applied to measure the relevance rate of the class d (i.e., $Rel^{e,d}$). It considers the mean value (i.e., $s-value$) that exist between two classes for enhancing the confidence rate. RFA-GA sets the value to 0, if it finds the any two groups are not extensively varied (i.e., $o-value$ 0.05) or the $s-value$ divided by $o-value$. Hence, the maximum value of $Rel^{e,d}$ indicates the relevancy level of the feature e to the class d. The part of features that having high relevance features are utilized in the formation of terminal set that are related to class d. This methodology not only performs the exclusion of features that are not relevant, but also helps in narrowing the search space which results in making the searching process one step more better.

Fig 1. Word Count

$$Rel^{s,d} = \begin{cases} 0 & if\ o- \\ s - value_{(class=d,\ class\neq d)} & ot \end{cases} \quad (8)$$

RFS-GA performs the classification in a varying number of aspects with 2 classes, namely optimistic-class and pessimistic-class. Initially, the objective of the two classes were different in developing the features. Then they utilizes the fitness function to show the performance differences. Finally, RFS-GA selects the features from the who available features to find the new features in the available search space.

IV. ABOUT MATLAB AND DATASET

MATLAB is a commercial software for computing the numerical operations used for scientific and engineering applications. It includes multiple inbuilt mathematical functions. It has a increased intensity towards all programming language by making them to access the advanced data structures, 2-dimension and 3-dimension graphical functions. The Amazon dataset [11] was used for experiments. It holds a collection of reviews from Amazon.com website. This dataset consist of three different types of files namely positive, negative and unlabelled in XML format. These files were made to extract using XML file splitter and reviews were converted into the text file. The dataset contains 1000 positive files and 1000 negative files for each domain. The reviews are about four item domains: Books (B), DVDs (D), Electronics (E) and Kitchen appliances (K) and are written in English language. For the experiment, labeled dataset of 1000 positive and 1000 negative files was used. An instance in each domain is recorded in Table 1. From this dataset, 12 cross-domain sentiment classification errands were constructed: $B \rightarrow D$; $B \rightarrow E$; $B \rightarrow K$; $D \rightarrow B$; $D \rightarrow E$; $D \rightarrow K$; $E \rightarrow B$; $E \rightarrow D$; $E \rightarrow K$; $K \rightarrow B$; $K \rightarrow D$; $K \rightarrow E$, where the word before arrow corresponds to the source domain and the word after an arrow corresponds to the target domain.

V.RESULTS AND DISCUSSION

A. Word Count

Fig 1 indicates the total number of words available in two domain specific and independent domain dataset, according to the cross domains. The corresponding values of Fig 1 is projected in Table 1

.

Table 1. Count of words in specific and independent domains

Source Domain to Cross Domain	Domain Specific Words	Domain Independent Words
$B \rightarrow D$	9744	11503
$B \rightarrow E$	4796	5250
$B \rightarrow K$	4325	4200
$D \rightarrow B$	9744	11130
$D \rightarrow E$	4781	5238
$D \rightarrow K$	4320	4205
$E \rightarrow B$	4769	16105
$E \rightarrow D$	4781	16466
$E \rightarrow K$	3723	4802
$K \rightarrow B$	4325	16549
$K \rightarrow D$	4320	16927
$K \rightarrow E$	3723	6296

B. Identification of Words

Fig 2 and Fig 3 represents the count of words identified by SentiWordNet [12] and RFS-GA in domain specific and independent domains datasets. The x-axis in Fig 2 and Fig 3 are plotted with cross domains and y-axis is plotted with word count. The results indicates that RFS-GA has identified more words in both the domains than the SentiWordNet [12].

The corresponding values of Fig 2 and Fig 3 is projected in Table 2

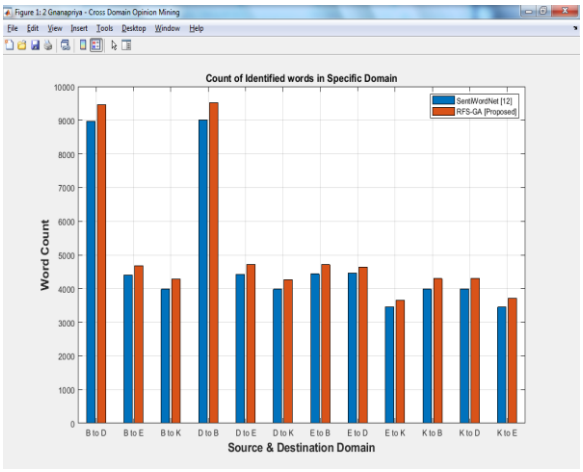

Fig 2. Identification of words in Domain Specific

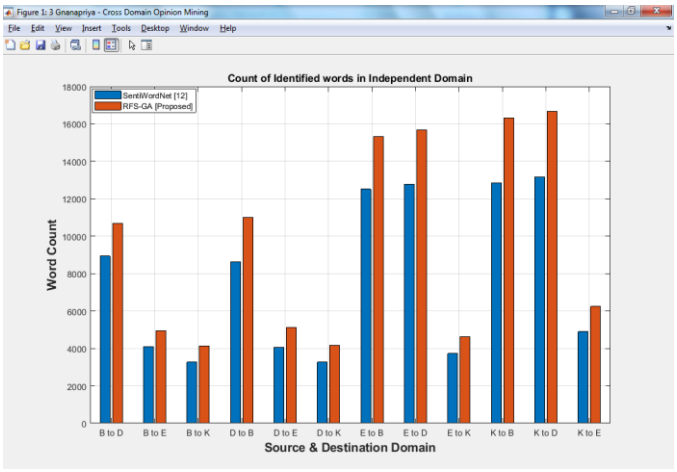

Fig 3. Identification of words in Independent Domain

Table 2. Count of identified words in domain specific and independent domains dataset

Source Domain to Cross Domain	Domain Specific Words		Domain Independent Words	
	SentiWordNet [12]	RFS-GA [Proposed]	SentiWordNet [12]	RFS-GA [Proposed]
$B \rightarrow D$	8972	9464	8934	10684
$B \rightarrow E$	4395	4680	4077	4931
$B \rightarrow K$	3979	4280	3262	4121
$D \rightarrow B$	9009	9516	8644	11008
$D \rightarrow E$	4418	4718	4068	5119
$D \rightarrow K$	3980	4263	3266	4160
$E \rightarrow B$	4430	4707	12508	15339
$E \rightarrow D$	4462	4640	12789	15674
$E \rightarrow K$	3454	3660	3729	4618
$K \rightarrow B$	3980	4296	12853	16311
$K \rightarrow D$	3987	4302	13147	16659
$K \rightarrow E$	3449	3713	4890	6240

C. Sensitivity Analysis

sensitivity analysis indicates the capacity to find out the correct cases. In order to estimate the sensitivity it is necessary to calculate the proportion of TP in the available cases. It is mathematically expressed as

$$Sensitivity = \frac{TP}{TP + FN} \quad (9)$$

Fig 4 and Fig 5 represents the sensitivity analysis of SentiWordNet [12] and RFS-GA in domain specific and independent domains datasets. The x-axis in Fig 4 and Fig 5 are plotted with cross domains and y-axis is plotted with sensitivity in percentage. The results shows that RFS-GA has better sensitivity than SentiWordNet [12] and it is because of better applying the crossover and mutation. The corresponding values of Fig 2 and Fig 3 is projected in Table 3.

.

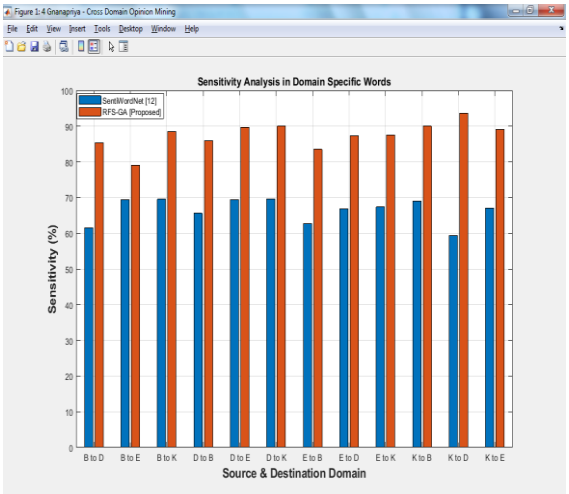

Fig 4. Sensitivity Analysis in Domain Specific

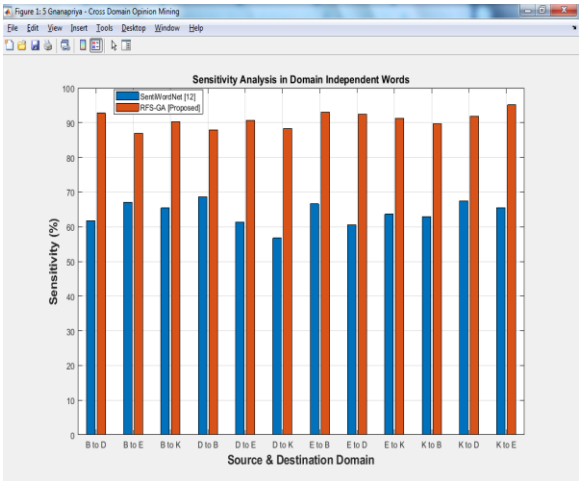

Fig 5. Sensitivity Analysis in Independent Domain

Table 3. Sensitivity Analysis in Specific and Independent Domains (in %)

Source Domain to Cross Domain	Domain Specific Words		Domain Independent Words	
	SentiWordNet [12]	RFS-GA [Proposed]	SentiWordNet [12]	RFS-GA [Proposed]
$B \rightarrow D$	61.47	85.39	61.81	92.73
$B \rightarrow E$	69.48	78.97	67.05	86.97
$B \rightarrow K$	69.54	88.47	65.39	90.33
$D \rightarrow B$	65.58	85.91	68.62	87.97
$D \rightarrow E$	69.32	89.63	61.27	90.62
$D \rightarrow K$	69.56	90.00	56.72	88.34
$E \rightarrow B$	62.69	83.49	66.54	92.98
$E \rightarrow D$	66.90	87.21	60.45	92.36
$E \rightarrow K$	67.35	87.47	63.62	91.29
$K \rightarrow B$	69.05	89.97	62.92	89.64
$K \rightarrow D$	59.41	93.67	67.48	91.78
$K \rightarrow E$	67.07	89.17	65.49	95.08

D. Specificity Analysis

Specificity analysis indicates the capacity to find out the strong classes correctly. In order to estimate the sensitivity it is necessary to calculate the proportion of TN in the available cases. It is mathematically expressed as

$$Specificity = \frac{TN}{TN + FP} \quad (10)$$

Fig 6 and Fig 7 represents the specificity analysis of SentiWordNet [12] and RFS-GA in domain specific and independent domains datasets. The x-axis in Fig 6 and Fig 7 are plotted with cross domains and y-axis is plotted specificity in percentage. The results shows that RFS-GA has better sensitivity than SentiWordNet [12] and it is because of better applying the crossover and mutation. The corresponding values of Fig 6 and Fig 7 is projected in Table 4.

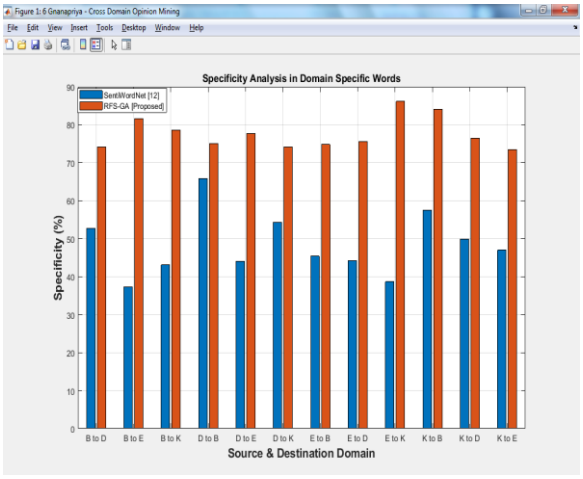

Fig 6. Specificity Analysis in Domain Specific

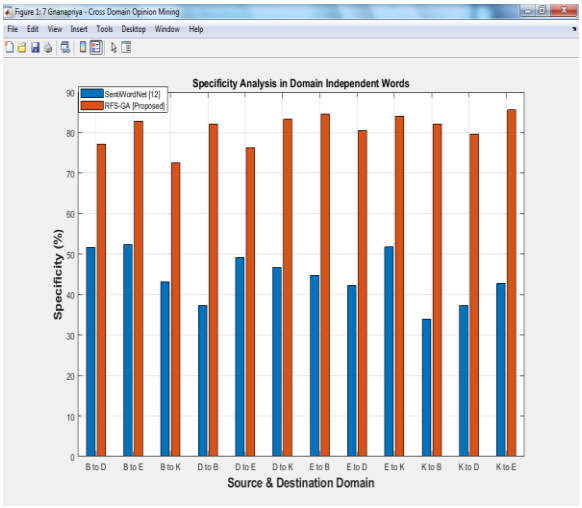

Fig 7. Specificity Analysis in Independent Domain

Table 4. Specificity Analysis in Specific and Independent Domains (in %)

Source Domain to Cross Domain	Domain Specific Words		Domain Independent Words	
	SentiWordNet [12]	RFS-GA [Proposed]	SentiWordNet [12]	RFS-GA [Proposed]
$B \rightarrow D$	52.78	74.12	51.73	77.15
$B \rightarrow E$	37.29	81.54	52.33	82.75
$B \rightarrow K$	43.07	78.60	43.19	72.54
$D \rightarrow B$	65.83	75.01	37.37	82.14
$D \rightarrow E$	44.00	77.62	49.24	76.22
$D \rightarrow K$	54.38	74.19	46.74	83.28
$E \rightarrow B$	45.37	74.81	44.79	84.56
$E \rightarrow D$	44.29	75.55	42.28	80.56
$E \rightarrow K$	38.65	86.11	51.77	84.02
$K \rightarrow B$	57.51	84.07	33.93	82.14
$K \rightarrow D$	49.82	76.44	37.37	79.58
$K \rightarrow E$	46.95	73.51	42.72	85.60

E. Precision Analysis

Precision is the division of related instances in the retrieved instances. In order to estimate the precision it is necessary to calculate the proportion from TP and FP. It is mathematically expressed as

$$Precision = \frac{TP}{TP + FP} \qquad (11)$$

Fig 8 and Fig 9 represents the precision analysis of SentiWordNet [12] and RFS-GA in domain specific and independent domains datasets. The x-axis in Fig 8 and Fig 9 are plotted with cross domains and y-axis is plotted precision in percentage. The results shows that RFS-GA has better precision than SentiWordNet [12] and it is because of better utilization of fitness function. The corresponding values of Fig 8 and Fig 9 is projected in Table 5.

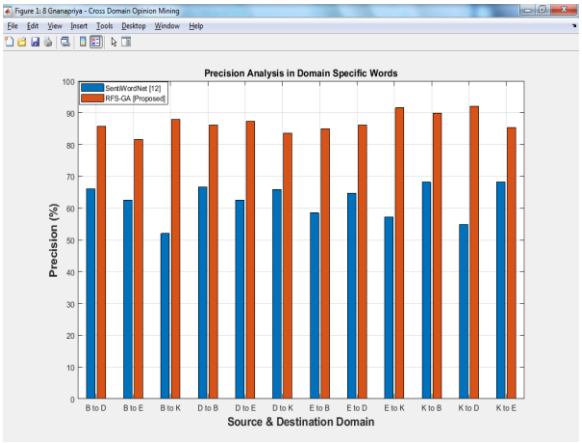

Fig 8. Precision Analysis in Domain Specific

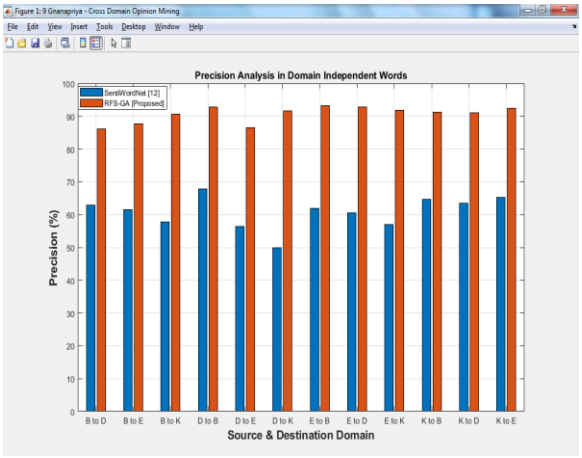

Fig 9. Precision Analysis in Independent Domain

Table 5. Precision Analysis in Specific and Independent Domains (in %)

Source Domain to Cross Domain	Domain Specific Words		Domain Independent Words	
	SentiWordNet [12]	RFS-GA [Proposed]	SentiWordNet [12]	RFS-GA [Proposed]
$B \rightarrow D$	65.99	85.71	62.91	86.17
$B \rightarrow E$	62.52	81.55	61.55	87.61
$B \rightarrow K$	52.02	87.86	57.76	90.72
$D \rightarrow B$	66.71	86.09	67.76	92.74
$D \rightarrow E$	62.52	87.29	56.34	86.58
$D \rightarrow K$	65.89	83.60	49.89	91.70
$E \rightarrow B$	58.50	84.98	61.90	93.22
$E \rightarrow D$	64.62	86.13	60.44	92.76
$E \rightarrow K$	57.28	91.54	57.00	91.86
$K \rightarrow B$	68.15	89.94	64.75	91.24
$K \rightarrow D$	54.87	91.94	63.50	90.99
$K \rightarrow E$	68.23	85.36	65.32	92.41

F. Recall Analysis

Recall is the part of related instances that are retrieved against the total available relevant instances. In order to estimate the Recall it is necessary to calculate the proportion from TN and FN. It is mathematically expressed as

$$Recall = \frac{TN}{TN + FN} \quad (12)$$

Fig 10 and Fig 11 represents the precision analysis of SentiWordNet [12] and RFS-GA in domain specific and independent domains datasets. The x-axis in Fig 10 and Fig 11 are plotted with cross domains and y-axis is plotted recall in percentage. The results shows that RFS-GA has better recall than SentiWordNet [12] and it is because of better selection of features. The corresponding values of Fig 10 and Fig 11 is projected in Table 6.Table 6-Recall Analysis in Specific and Independent Domains (in %)

Source Domain to Cross Domain	Domain Specific Words		Domain Independent Words	
	SentiWordNet [12]	RFS-GA [Proposed]	SentiWordNet [12]	RFS-GA [Proposed]
$B \rightarrow D$	61.47	85.39	61.81	92.73
$B \rightarrow E$	69.48	78.97	67.05	86.97
$B \rightarrow K$	69.54	88.47	65.39	90.33
$D \rightarrow B$	65.58	85.91	68.62	87.97
$D \rightarrow E$	69.32	89.63	61.27	90.62
$D \rightarrow K$	69.56	90.00	56.72	88.34
$E \rightarrow B$	62.69	83.49	66.54	92.98
$E \rightarrow D$	66.90	87.21	60.45	92.36
$E \rightarrow K$	67.35	87.47	63.62	91.29
$K \rightarrow B$	69.05	89.97	62.92	89.64
$K \rightarrow D$	59.41	93.67	67.48	91.78
$K \rightarrow E$	67.07	89.17	65.49	95.08

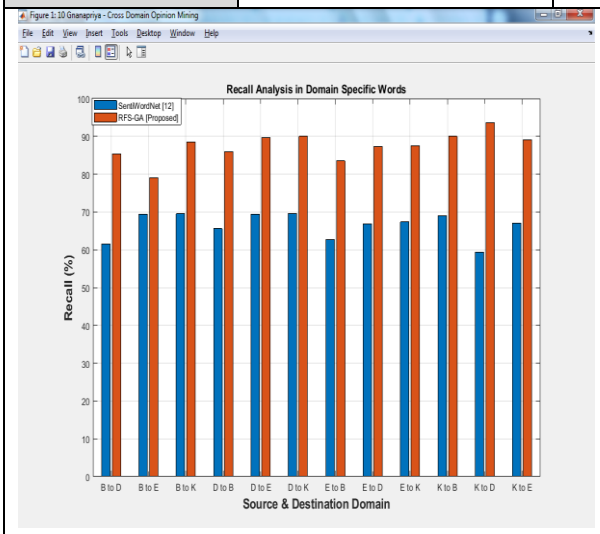

Fig 10. Recall Analysis in Domain Specific

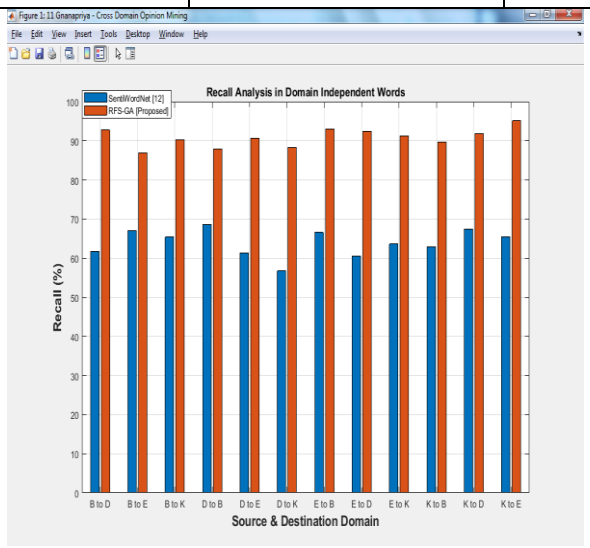

Fig 11. Recall Analysis in Independent Domain

G. Accuracy Analysis

Accuracy indicates the algorithms ability to make differentiation among the better classes and correct classes. In order to estimate the accuracy it is necessary to calculate the proportion from TP, TN, FP and FN. It is mathematically expressed as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

Fig 12 and Fig 13 represents the accuracy analysis of SentiWordNet [12] and RFS-GA in domain specific and independent domains datasets. The x-axis in Fig 12 and Fig 13 are plotted with cross domains and y-axis is plotted with accuracy in percentage. The results shows that RFS-GA has better accuracy than SentiWordNet [12] and it is because of better selection and fitness function applicability. The corresponding values of Fig 10 and Fig 11 is projected in Table 7.

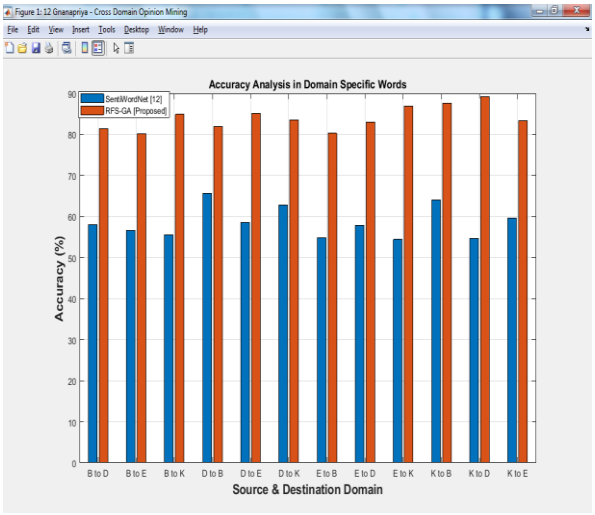

Fig 12. Accuracy Analysis in Domain Specific

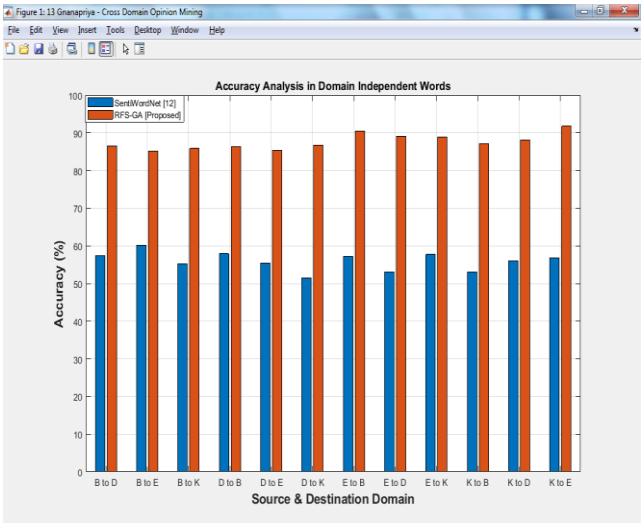

Fig 13. Accuracy Analysis in Independent Domain

Table 7. Accuracy Analysis in Specific and Independent Domains (in %)

Source Domain to Cross Domain	Domain Specific Words		Domain Independent Words	
	SentiWordNet [12]	RFS-GA [Proposed]	SentiWordNet [12]	RFS-GA [Proposed]
$B \rightarrow D$	57.98	81.39	57.48	86.59
$B \rightarrow E$	56.63	80.24	60.17	85.22
$B \rightarrow K$	55.52	84.88	55.24	85.85
$D \rightarrow B$	65.70	82.02	57.91	86.35
$D \rightarrow E$	58.53	85.21	55.46	85.27
$D \rightarrow K$	62.86	83.58	51.56	86.71
$E \rightarrow B$	54.92	80.28	57.28	90.42
$E \rightarrow D$	57.93	82.95	53.06	89.17
$E \rightarrow K$	54.43	86.97	57.71	88.85
$K \rightarrow B$	64.07	87.69	53.02	87.20
$K \rightarrow D$	54.68	89.21	55.97	88.02
$K \rightarrow E$	59.61	83.44	56.89	91.75

H. F-Measure Analysis

F-Measure is utilized to measure how far the accuracy is correct. It is also describes as the weighted harmonic mean of precision and recall. In order to estimate the f-measure it is necessary to calculate the proportion from precision and recall. It is mathematically expressed as

$$F - Measure = 2 \times \left(\frac{Precision}{Precision} \right. \quad (14)$$

Fig 14 and Fig 15 represents the f-measure analysis of SentiWordNet [12] and RFS-GA in domain specific and independent domains datasets. The x-axis in Fig 14 and Fig 15 are plotted with cross domains and y-axis is plotted with f-measure in percentage. The results shows that RFS-GA has better f-measure than SentiWordNet [12] and it is because of better selection and fitness function applicability. The corresponding values of Fig 14 and Fig 15 is projected in Table 8.

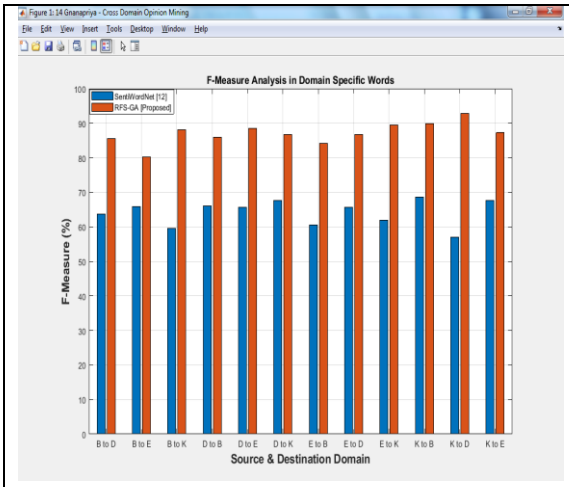

Fig 14. F-Measure Analysis in Domain Specific

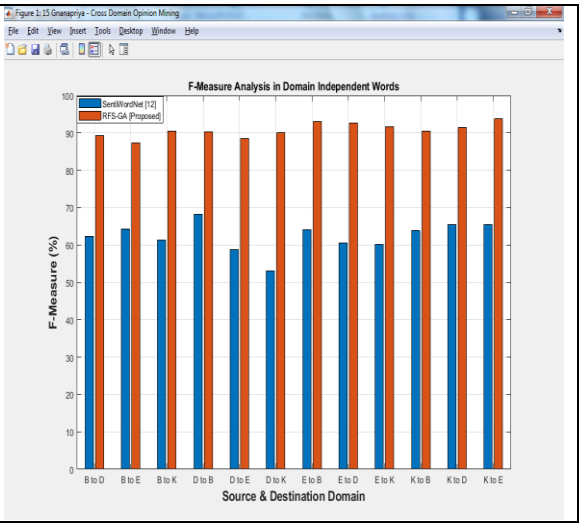

Fig 15. F-Measure Analysis in Independent Domain

Table 8. F-Measure Analysis in Specific and Independent Domains (in %)

Source Domain to Cross Domain	Domain Specific Words		Domain Independent Words	
	SentiWordNet [12]	RFS-GA [Proposed]	SentiWordNet [12]	RFS-GA [Proposed]
$B \rightarrow D$	63.65	85.55	62.36	89.33
$B \rightarrow E$	65.82	80.24	64.18	87.29
$B \rightarrow K$	59.52	88.17	61.33	90.52
$D \rightarrow B$	66.14	86.00	68.19	90.29
$D \rightarrow E$	65.74	88.44	58.71	88.55
$D \rightarrow K$	67.67	86.68	53.08	89.99
$E \rightarrow B$	60.53	84.23	64.13	93.10
$E \rightarrow D$	65.74	86.66	60.45	92.56
$E \rightarrow K$	61.91	89.46	60.13	91.58
$K \rightarrow B$	68.60	89.95	63.82	90.44
$K \rightarrow D$	57.05	92.80	65.43	91.38
$K \rightarrow E$	67.64	87.22	65.41	93.73

VI. CONCLUSION

Multiple proposals were made for opinion mining but it didn't solve the issues for cross domain. In this article robustious feature selection based genetic algorithm is proposed to solve the problems that arise in cross domain opinion mining. The proposed algorithm initially performs the feature selection in a better and wisely applies the genetic algorithm in optimistic class and pessimistic class. Due to this the proposed algorithm has attained the better results in all the considered benchmark performance metrics than the previous work.

REFERENCES:

1. Bissan. G., Joe.N., "High dimensional data classification and feature selection using support vector machines," European Journal of Operational Research, Volume 265, Issue 3,Pages 993-1004,2018.
2. Iti. C., Erik. C., Roy. E. W., Francisco. H.,"Distinguishing between facts and opinions for sentiment analysis: Survey and challenges," Information Fusion, Volume 44,Pages 65-77,2018.
3. Aitor. G., Montse. C., German. R.,"W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis," Expert Systems with Applications, Volume 91,Pages 127-137, 2018.
4. Aytuğ. O., Serdar. K., Hasan. B.,"A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," Expert Systems with Applications, Volume 62,Pages 1-16, 2016.
5. Mu-Yen.C., Ting-Hsuan. C.,"Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena," Future Generation Computer Systems, Volume 96, Pages 692-699, 2019.
6. Farhan. H. K., Usman.Q., Saba. B.,"SWIMS: Semi-supervised subjective feature weighting and intelligent model selection for sentiment analysis," Knowledge-Based Systems, Volume 100, Pages 97-111, 2016.
7. Alireza.Y., Roliana. I., Haza.N. A. H.,"Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis," Expert Systems with Applications, Volume 75, Pages 80-93, 2017.
8. Tomáš. K., Michal. N., Jiří. P.,"Improving sentiment analysis performance on morphologically rich languages: Language and domain independent approach,"Computer Speech & Language,Volume 56,Pages 36-51,2019.

9. Claudia. D., Alex.M., Domenico. P., Emanuele. S. "Social information discovery enhanced by sentiment analysis techniques," Future Generation Computer Systems, Volume 95, Pages 816-828, 2019.
10. Nádia. F.F., Eduardo. R. H., Estevam. R. H. "Tweet sentiment analysis with classifier ensembles," Decision Support Systems, Volume 66, Pages 170-179, 2014.
11. John. B., Mark. D., Fernando. P., "Biographies, Bollywood, boomboxes and blenders: domain adaptation for sentiment classification" 45th Annual Meeting of the Computational Linguistics (ACL), Prague, Czech Republic, Pages 440-447, 2007.
12. Baccianella. S., Esuli. A., Sebastiani. F. "SentiWordNet 3.0: An Enhance Lexical Resource for Sentiment Analysis and Opinion Mining," 7th Language Resources and Evaluation Conference (LREC 2010), Valletta, Malta, Pages 2200-2204, 2010.