ILS Z-534 Information Retrieval Theory and Practice

Final Project Report

Text Classification on Yelp Dataset

Naveen Shanmugasundaram – 0003400672

Niranjan Pachaiyappan - 0003394807

Venkatesh Thirupathisamy - 0003395917

Vignesh Vijayaraghavan - 0003393011

14/12/2014

Abstract

The goal of this project is to classify set of text information from the “Tips” and “Review” given by a particular user in Yelp Website on whether they recommend a particular business or not. “Tip” information is useful for Business owners to improve on their business and users who do not want to read large reviews. “Review” information is useful for gaining knowledge about the service of a particular business and user experience.

The Project is divided into two tasks.

Task-1: Classifying Categories of a particular business from the user tips and reviews. In the first and second approach for task - 1 businesses are categorized by unigram, bigram and trigram classification of “TIPS” with a comparative analysis of performance with and without information gain for attribute selection is done. Used Comparative analysis of two machine learning algorithms (**Multinomial Naïve Bayes and Support Vector Machines**) with the help of WEKA Machine Learning Package for this approach.

In the third approach for task - 1 businesses are categorized by unigram, bigram and trigram classification of both ” TIP ” and “ REVIEWS ” with information gain as the attribute selection criteria. Used machine learning algorithm **Support Vector Machines** with the help of WEKA Machine Learning Package for this approach.

Task -2: Predicting whether a particular review gives High Rating or Low Rating based on the response of words in the review text. In the first approach for task - 2 reviews are categorized by unigram, bigram and trigram classification with a comparative analysis of performance as information gain is used for attribute selection. Used Comparative analysis of two machine learning algorithms (**Multinomial Naïve Bayes and Support Vector Machines**) with the help of WEKA Machine Learning Package.

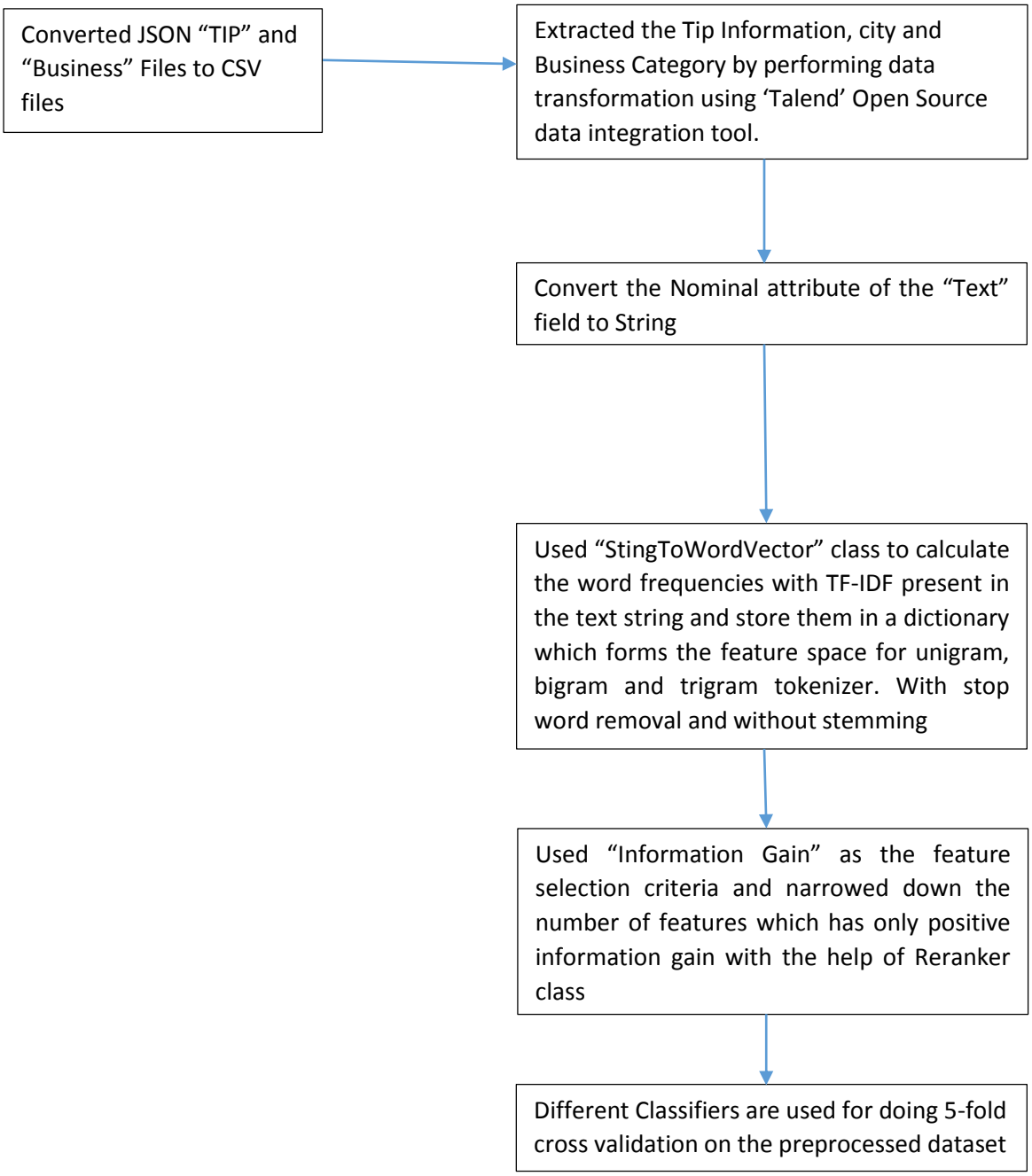
In the second approach for task - 2 reviews are categorized by N-gram (subset of unigrams, bigrams and trigrams) classification with information gain as the attribute selection criteria. Used Comparative analysis of two machine learning algorithms (**Multinomial Naïve Bayes and Support Vector Machines**) with the help of WEKA Machine Learning Package for this approach.

.

Methodology

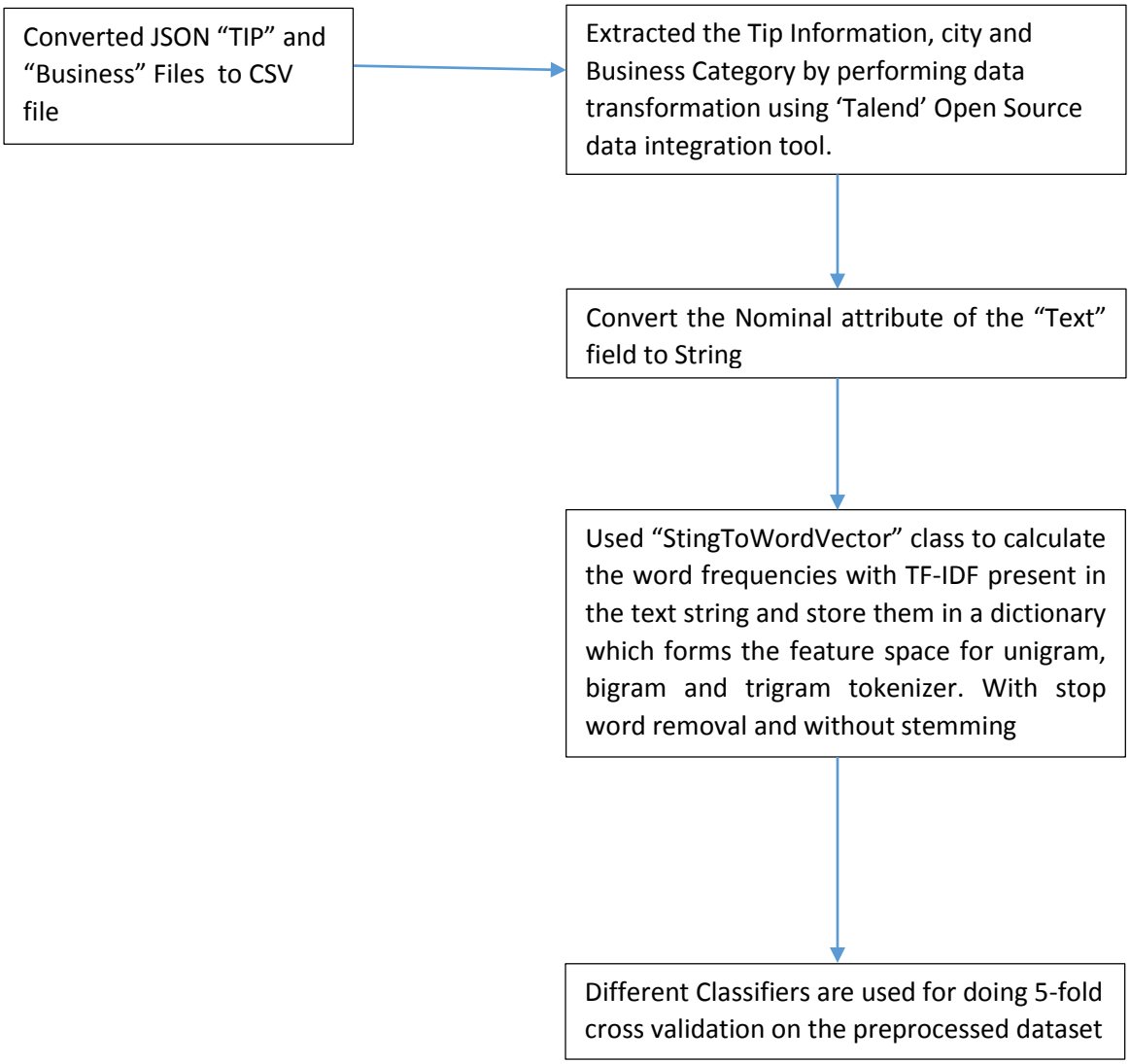
TASK-1:

The Following Methodology is used for performing the Text Classification for Task-1 -> Approach -1



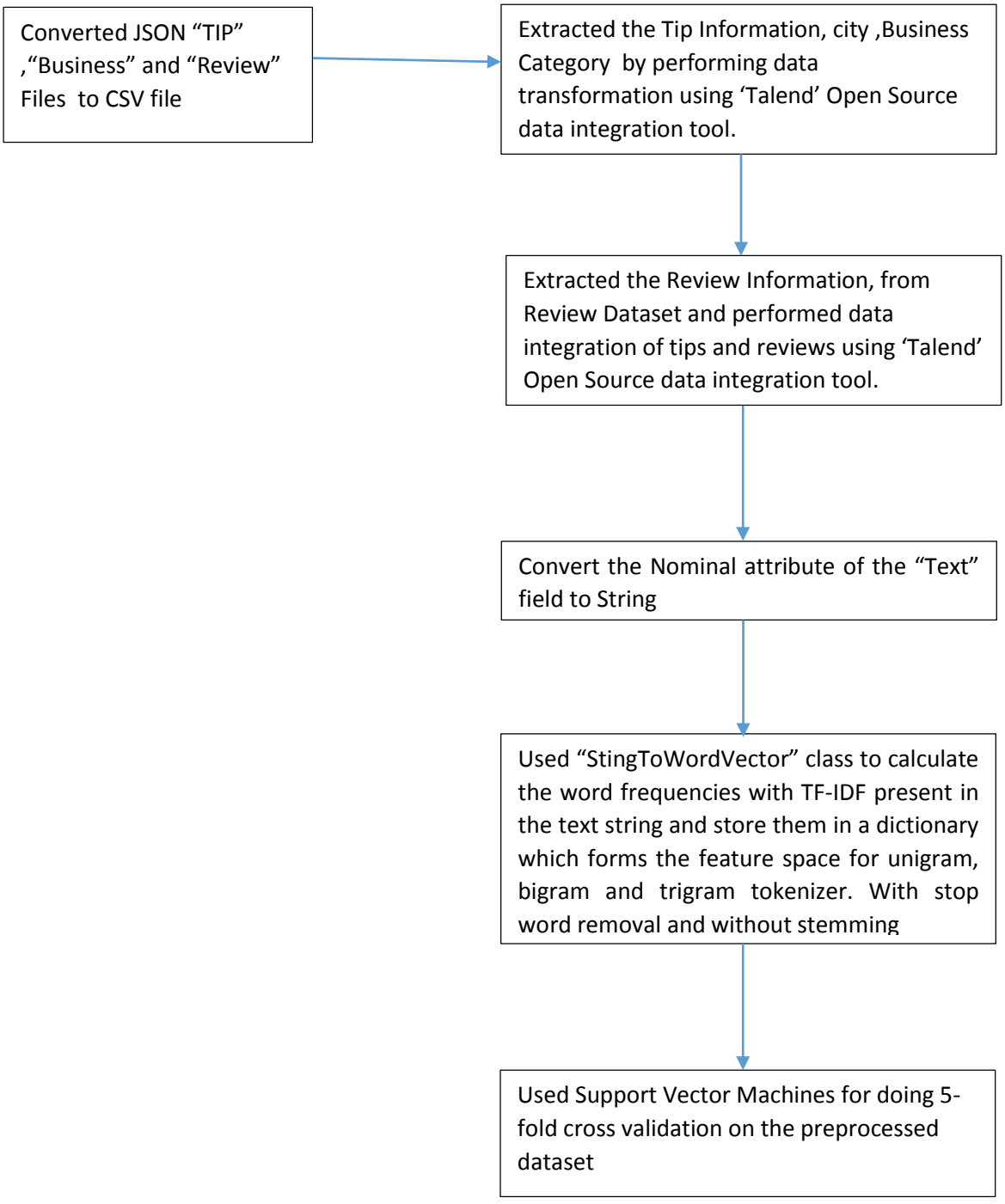
TASK-1:

The Following Methodology is used for performing the Text Classification for Task-1 -> Approach -2



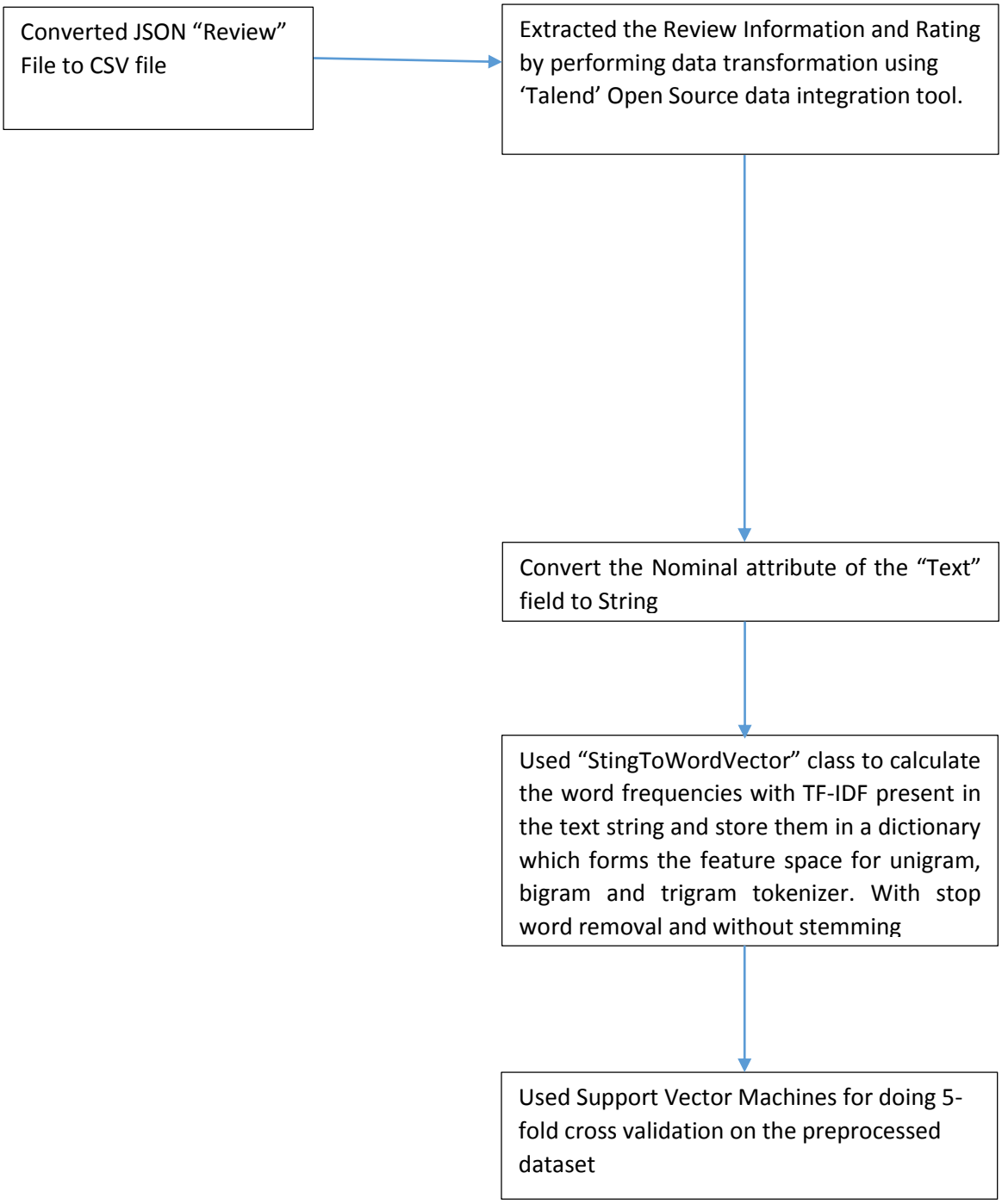
TASK-1:

The Following Methodology is used for performing the Text Classification for Task-1 -> Approach -3



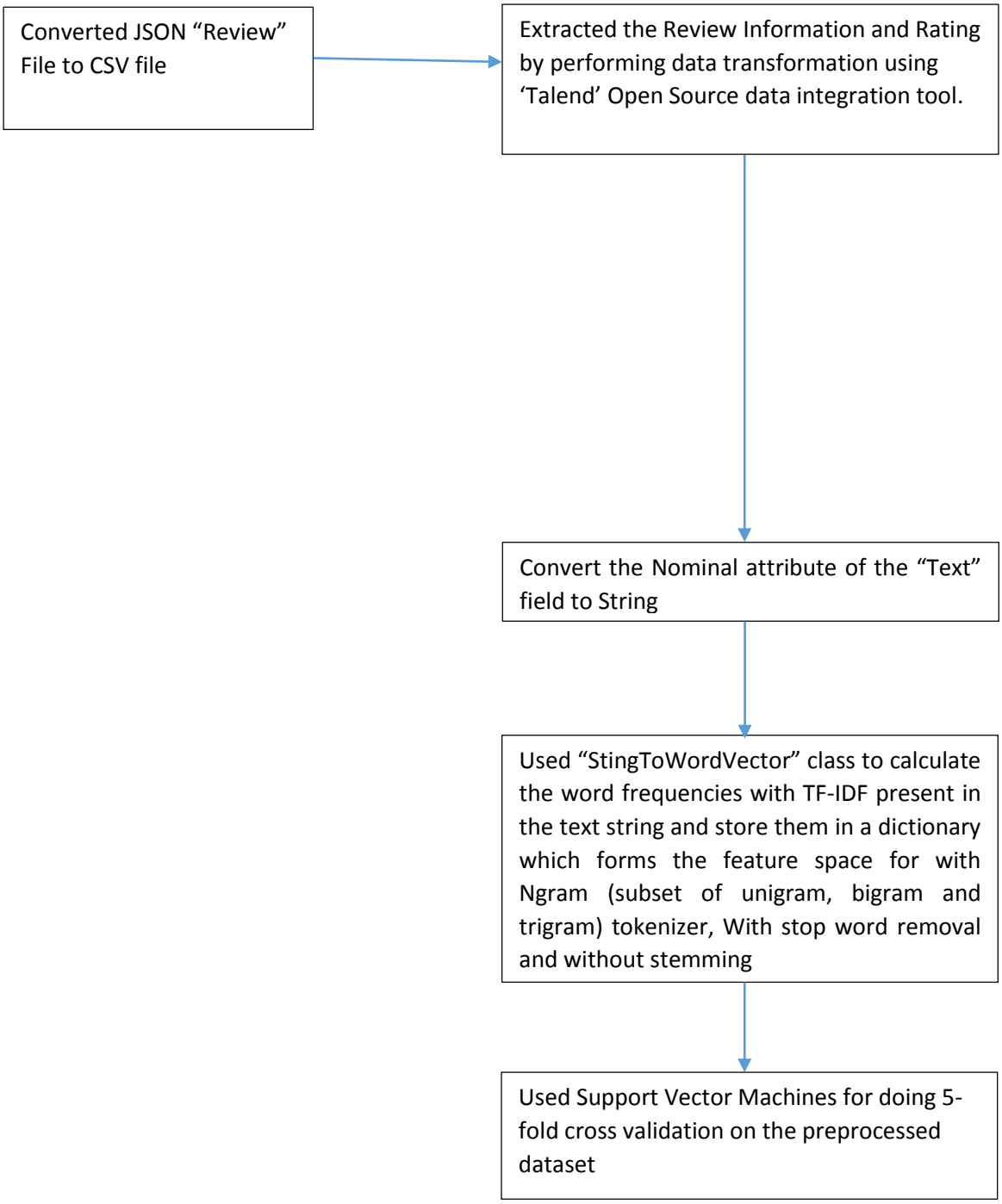
TASK-2:

The Following Methodology is used for performing the Text Classification for Task-2 -> Approach -1

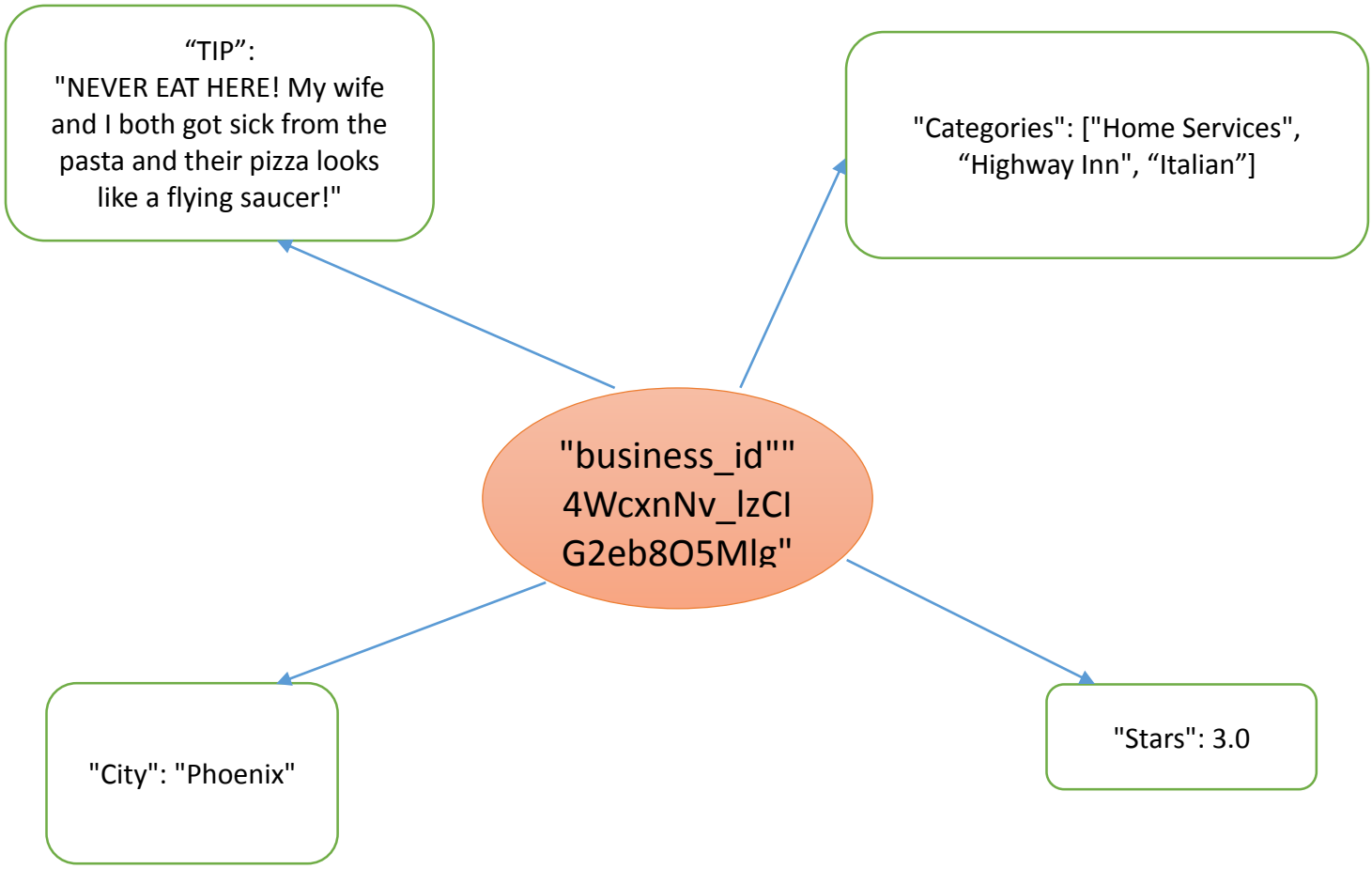


TASK-2:

The Following Methodology is used for performing the Text Classification for Task-2 -> Approach -2



Data Set Description



(Number of Categories =15)

Task-1 -> Approach -1 (Tip with Information Gain) :

Approach / Count	**UNIGRAM**	**BIGRAM**	**TRIGRAM**
Total Number of Instances	7217	7217	7217
Number of Features(after information gain selection)	1722	1014	252

Task-1 -> Approach -2 (Tip without Information Gain):

(Number of Categories =15)

Approach / Count	UNIGRAM	BIGRAM	TRIGRAM
Total Number of Instances	7217	7217	7217
Number of Features(with no information gain selection)	10761	20318	70300

Task-1 -> Approach -3 (Tip and Review with Information Gain):

(Number of Categories =16)

Approach / Count	UNIGRAM	BIGRAM	TRIGRAM
Total Number of Instances	1743	1743	1743
Number of Features(with no information gain selection)	1912	3099	1542

Task -2 -> Approach -1(Review with Information Gain individual grams):

Approach / Count	UNIGRAM	BIGRAM	TRIGRAM
Total Number of Instances	27844	27844	27844
Number of Features(with no information gain selection)	142	43	54

Task -2 -> Approach -2(Review with Information Gain with subset of N-grams):

Approach / Count	N-GRAM
Total Number of Instances	27844
Number of Features(with no information gain selection)	115

Results and Evaluation Metrics

Task-1- Approach -1:

Unigram

Multinomial Naïve Bayes

Results	Values
Accuracy	78.3982 %
Precision	0.823
Recall	0.784
False Positive Rate	0.784
True Positive Rate	0.047

Support Vector Machine

Results	Values
Accuracy	76.7632 %
Precision	0.779
Recall	0. 768
False Positive Rate	0.208
True Positive Rate	0.768

Task-1- Approach -1:

Bigram

Multinomial Naïve Bayes

Results	Values
Accuracy	70.0568 %
Precision	0.693
Recall	0.701
False Positive Rate	0.148
True Positive Rate	0.701

Support Vector Machine

Results	Values
Accuracy	61.2581 %
Precision	0.669
Recall	0. 613
False Positive Rate	0.374
True Positive Rate	0.613

Task-1- Approach -1:

Trigram

Multinomial Naïve Bayes

Results	Values
Accuracy	57.711 %
Precision	0.576
Recall	0.577
False Positive Rate	0.360
True Positive Rate	0.577

Support Vector Machine

Results	Values
Accuracy	54.8012 %
Precision	0.631
Recall	0. 548
False Positive Rate	0.439
True Positive Rate	0.548

Task-1- Approach -2:

Unigram

Multinomial Naïve Bayes

Results	Values
Accuracy	75.1559
Precision	0.753
Recall	0.752
False Positive Rate	0.009
True Positive Rate	0.752

Support Vector Machine

Results	Values
Accuracy	56.5045
Precision	0.674
Recall	0.565
False Positive Rate	0.426
True Positive Rate	0.565

Task-1- Approach -2:

Bigram

Multinomial Naïve Bayes

Results	Values
Accuracy	75.1559
Precision	0.753
Recall	0.752
False Positive Rate	0.009
True Positive Rate	0.752

Support Vector Machine

Results	Values
Accuracy	49.4804%
Precision	0.245
Recall	0. 495
False Positive Rate	0.495
True Positive Rate	0.495

Task-1- Approach -3:

Unigram

Support Vector Machine

Results	Values
Accuracy	92.8285
Precision	0.920
Recall	0.928
False Positive Rate	0.023
True Positive Rate	0.928

Task-1- Approach -3:

Bigram

Support Vector Machine

Results	Values
Accuracy	86.288
Precision	0.856
Recall	0.863
False Positive Rate	0.062
True Positive Rate	0.083

Task-1- Approach -3:

Trigram

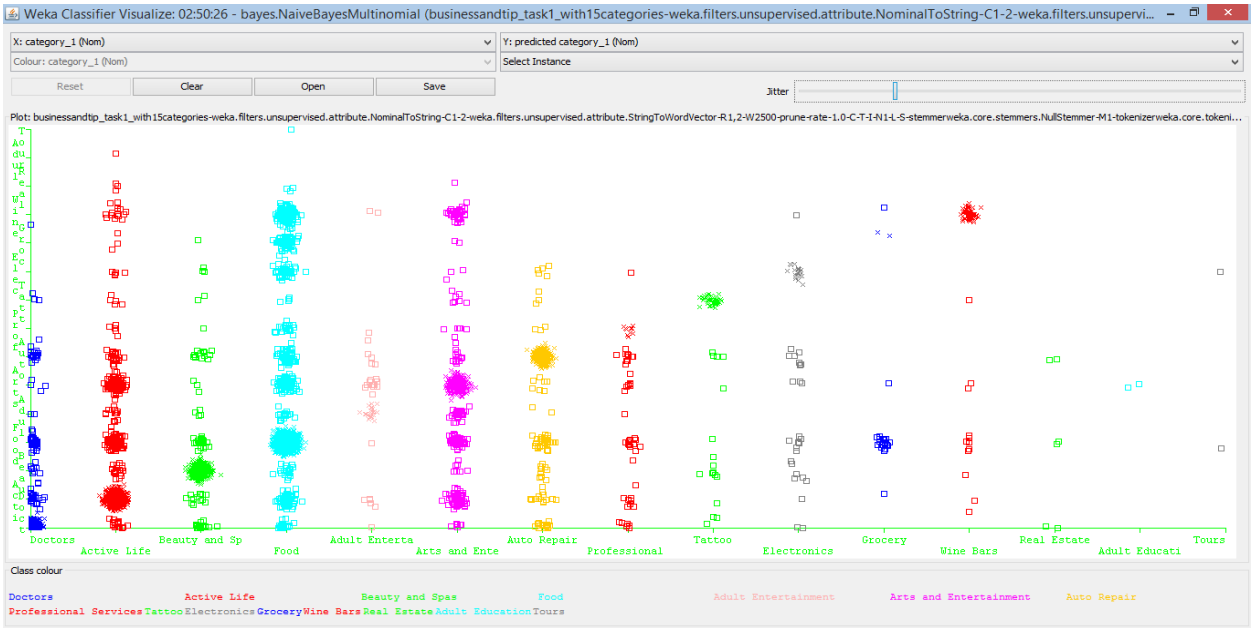
Support Vector Machine

Results	Values
Accuracy	74.2398
Precision	0.762
Recall	0.742
False Positive Rate	0.126
True Positive Rate	0.742

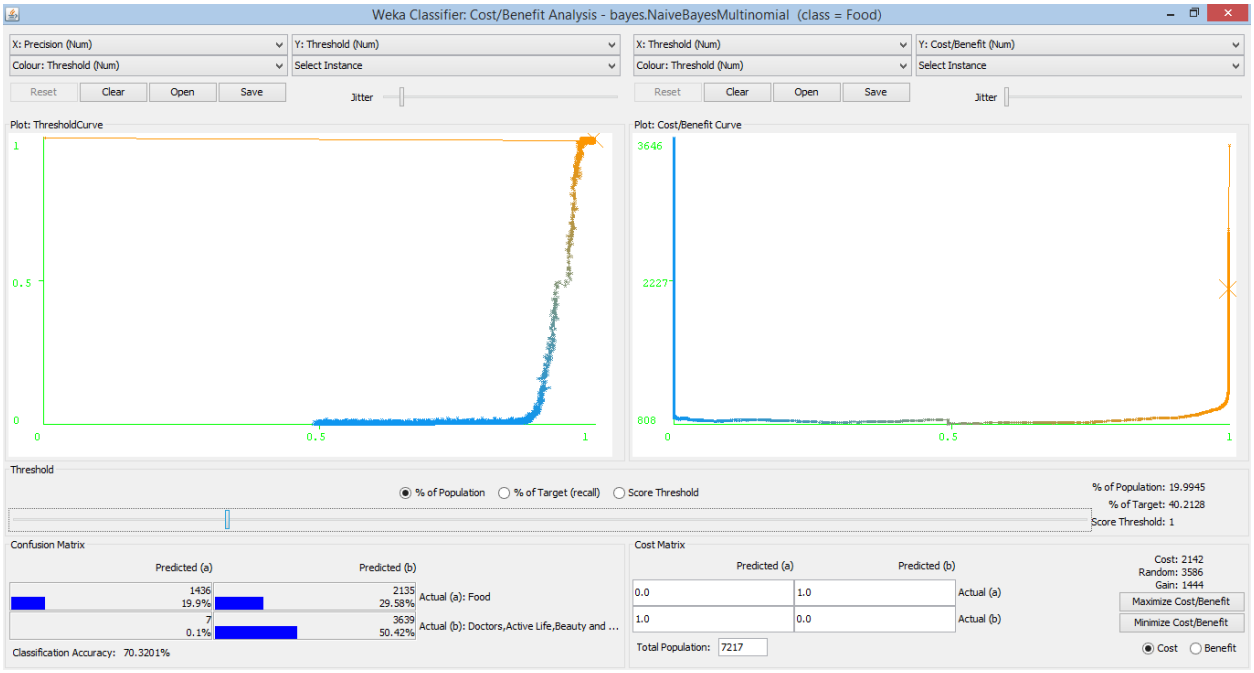
Graphical Representations of Classification Errors, Precision and Cost Benefit Curves for the classifiers(more result images are uploaded to git repository)

Task-1- Approach -1: (Unigram)

Multinomial Naïve Bayes



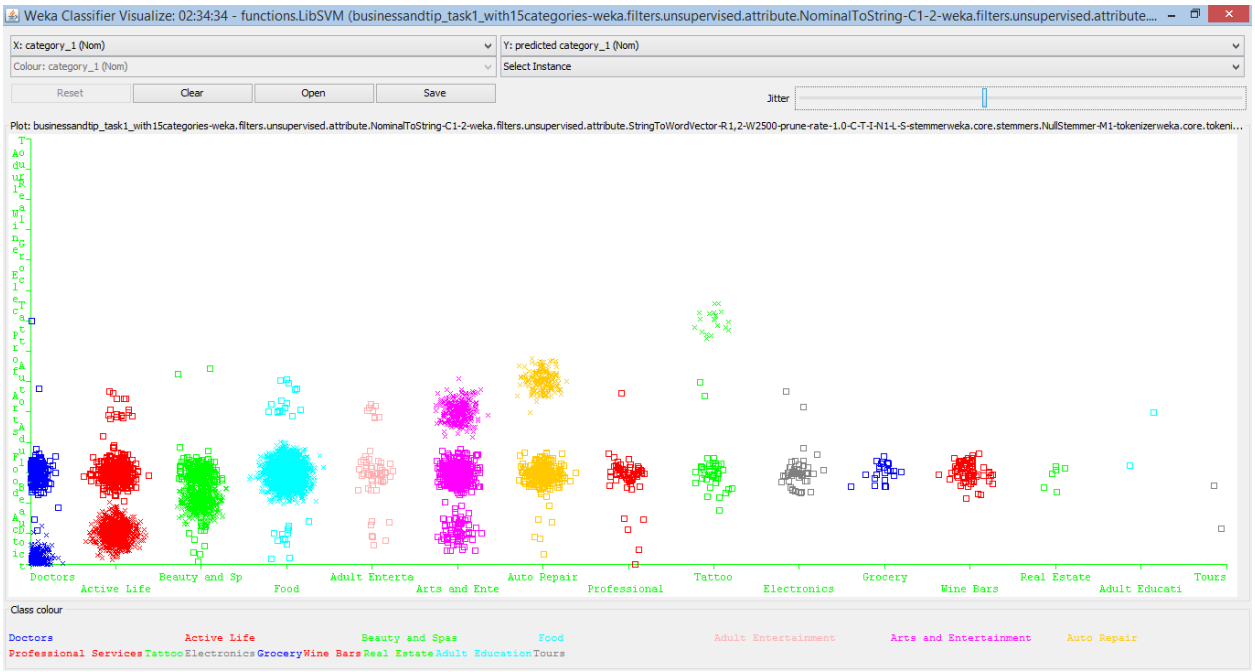
Classification Error:



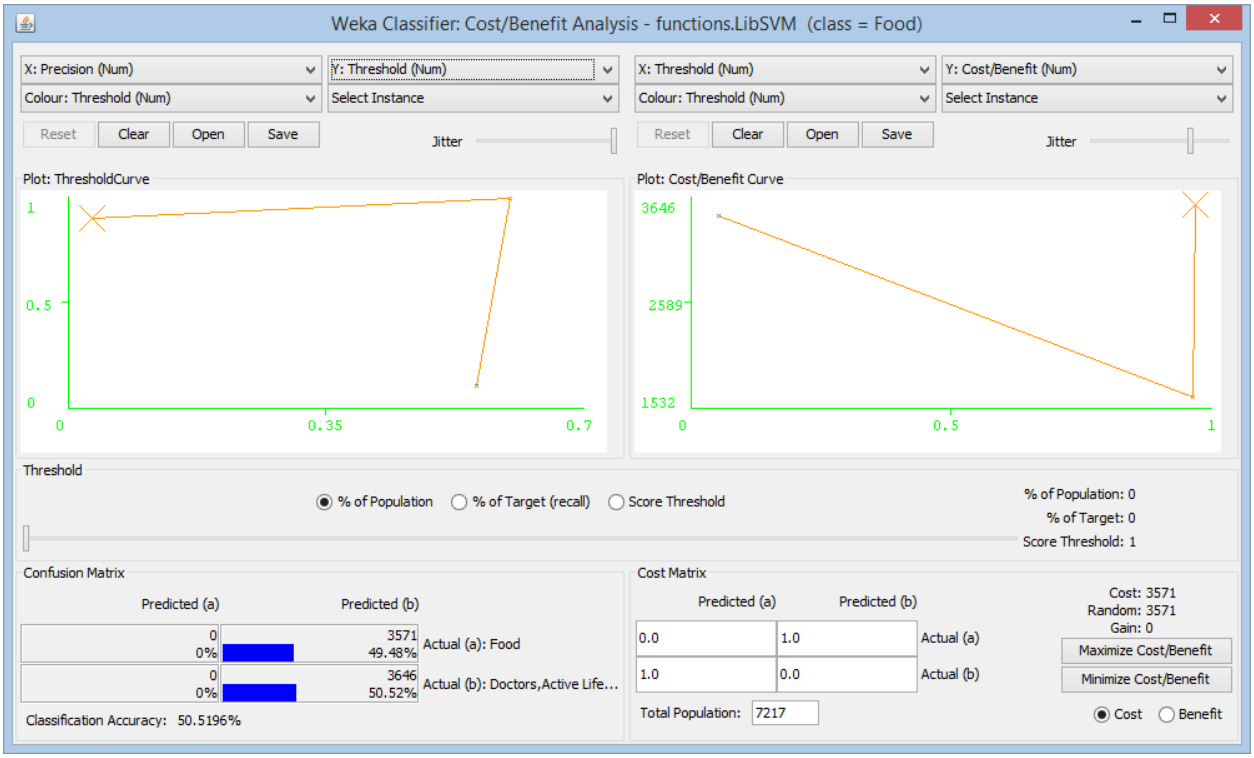
Cost Benefit Curve : (For FOOD Category)

SVM's

Classification Error:



Cost Benefit Curve: (For FOOD Category)

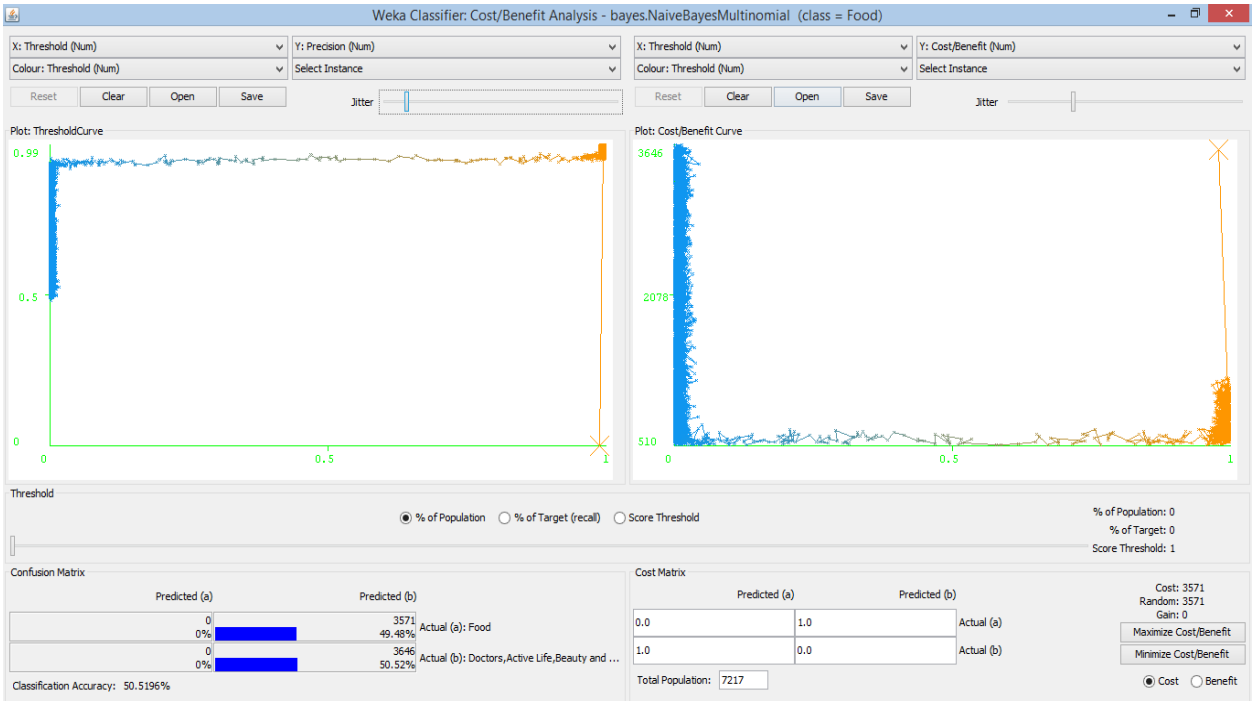


Task-1- Approach -2: (Unigram)

Multinomial Naïve Bayes

Classification Error:

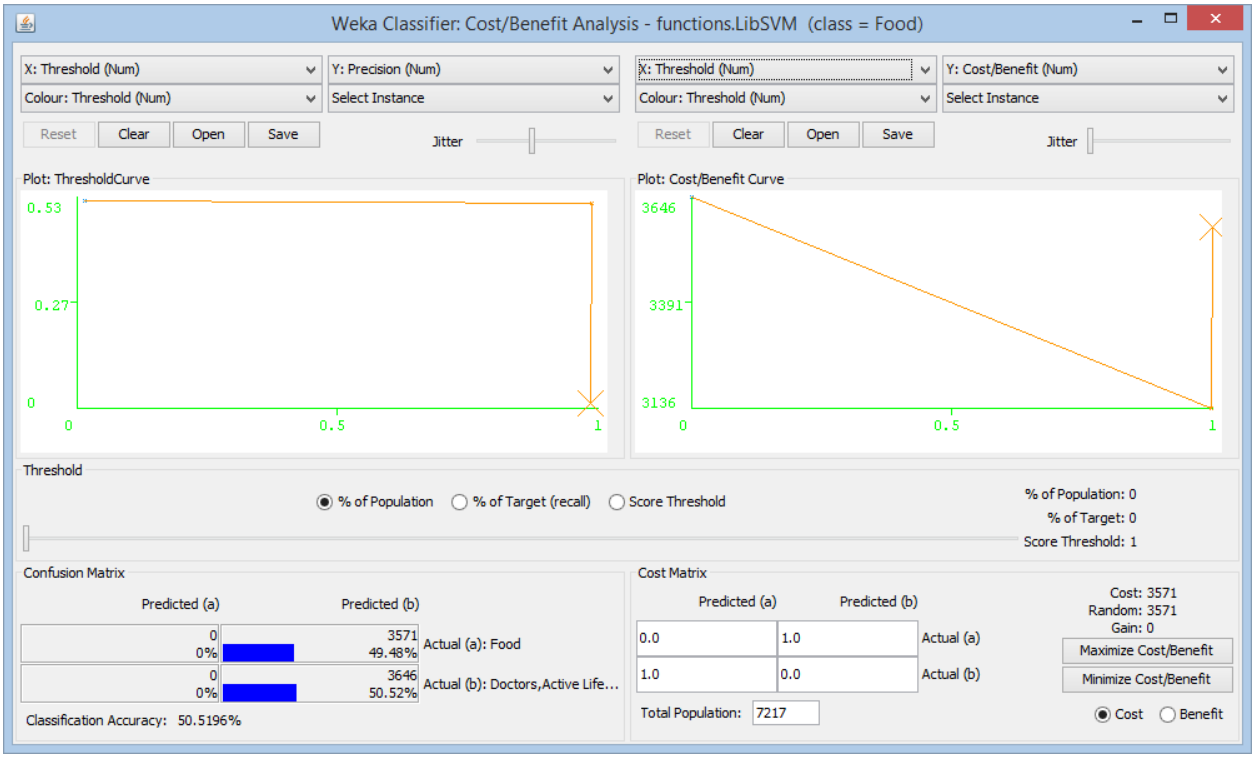
Cost Benefit Curve: (For FOOD Category)



SVM's

Classification Error:

Cost Benefit Curve: (For FOOD Category)

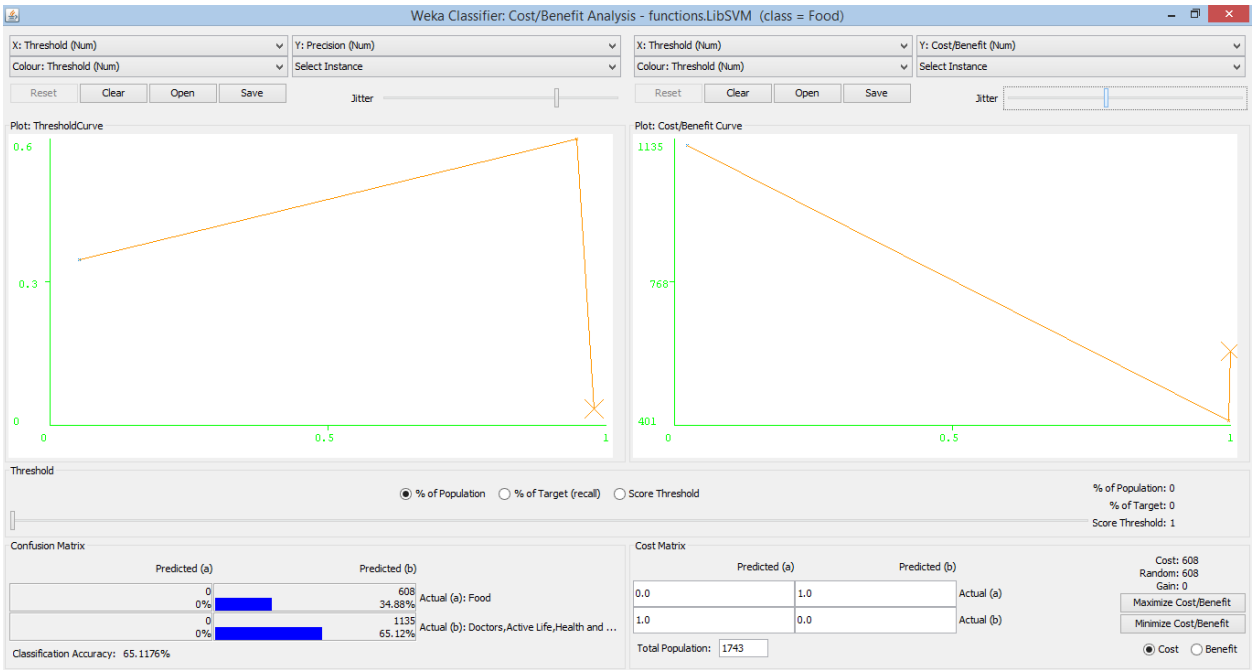


Task-1- Approach -3: (Unigram)

SVM's

Classification Error:

Cost Benefit Curve: (For FOOD Category)

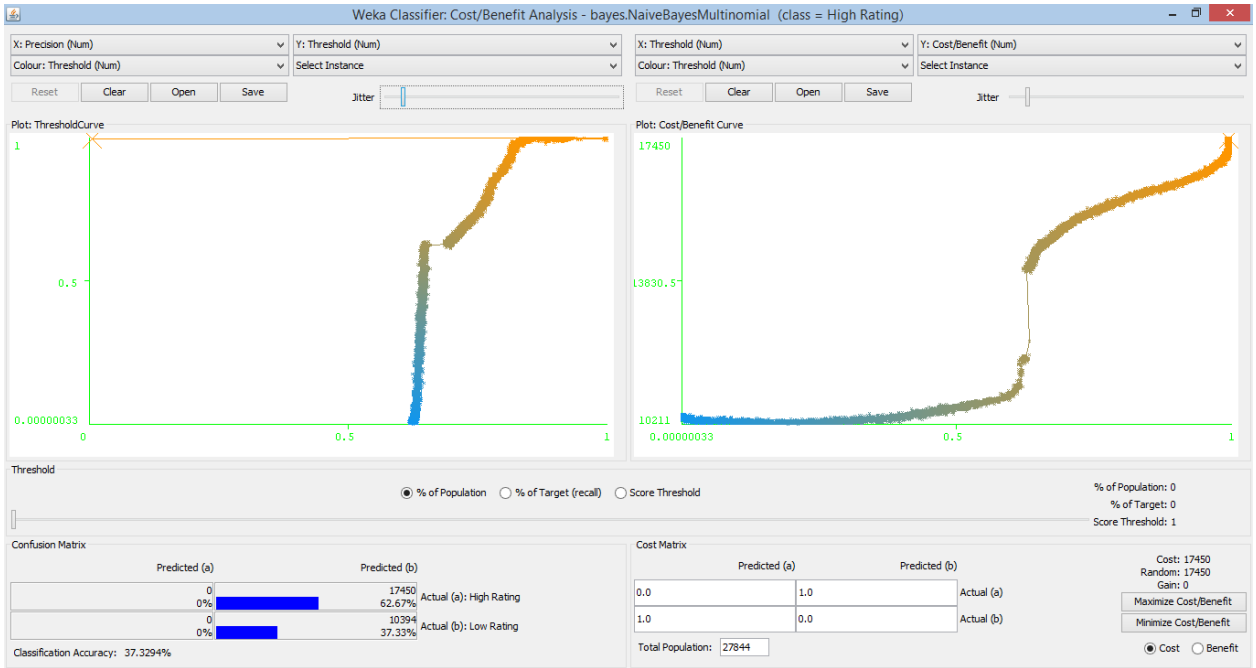


Task-2- Approach -1: (Unigram)

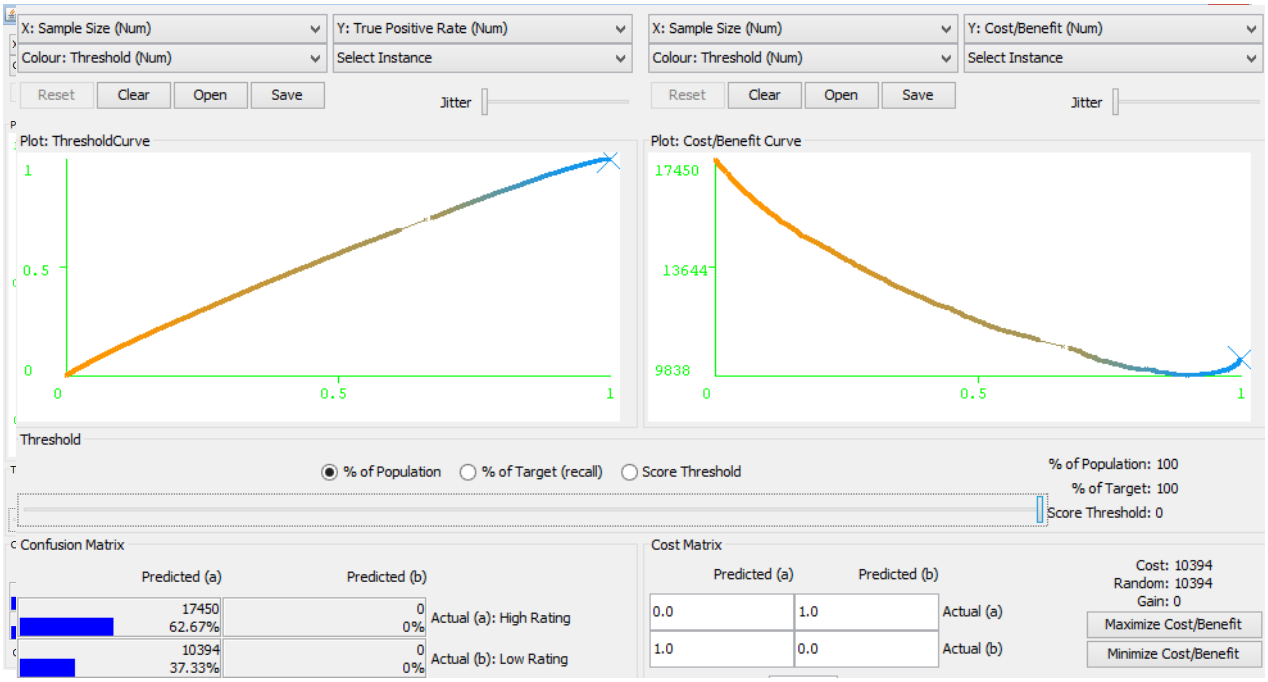
Multinomial Naïve Bayes

Classification Error:

Cost Benefit Curve: (For High rating)

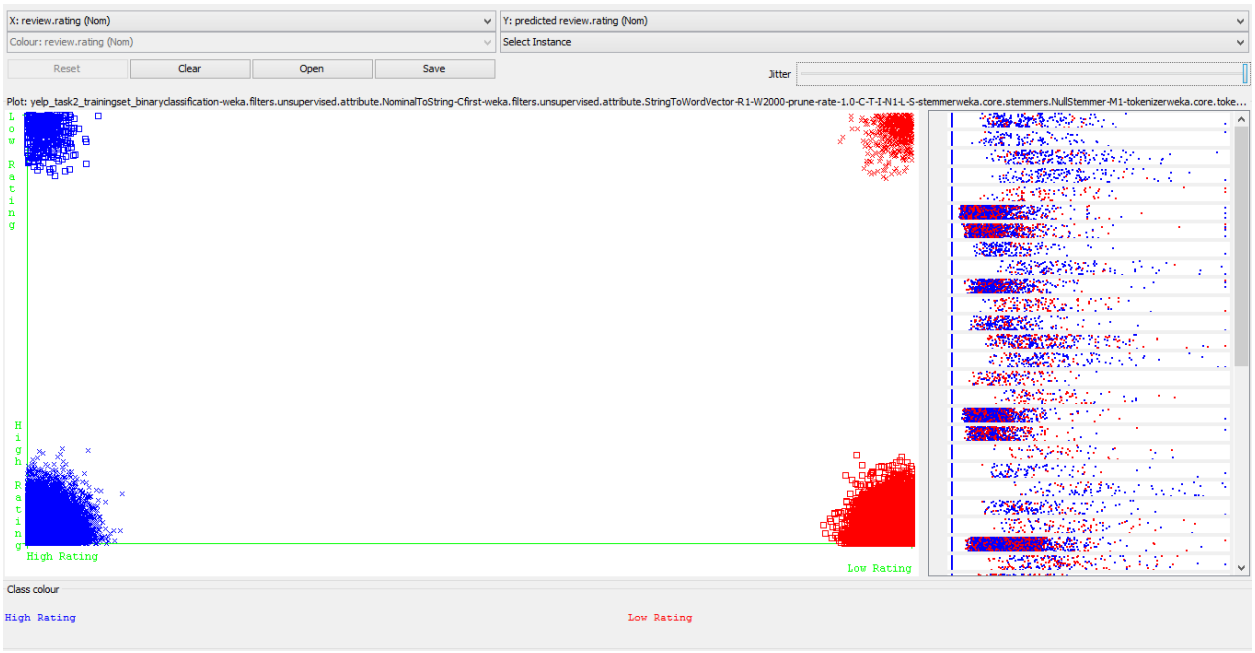


Cost Benefit Curve: (For Low rating)

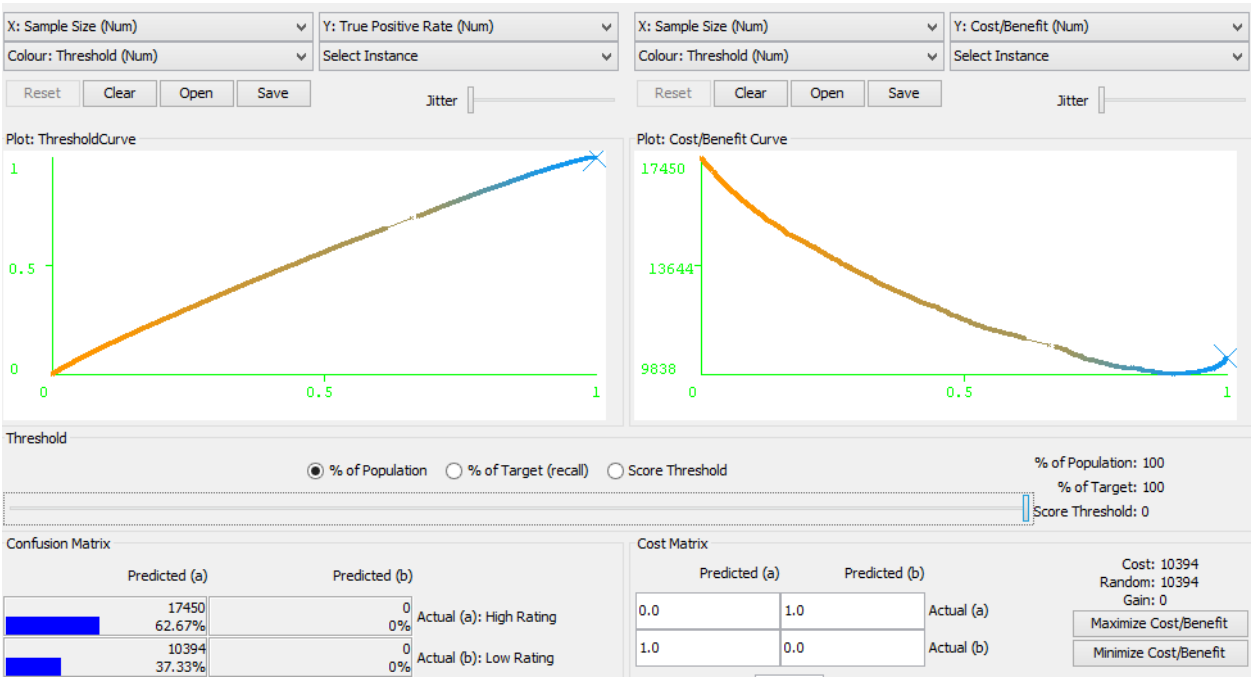


SVM's

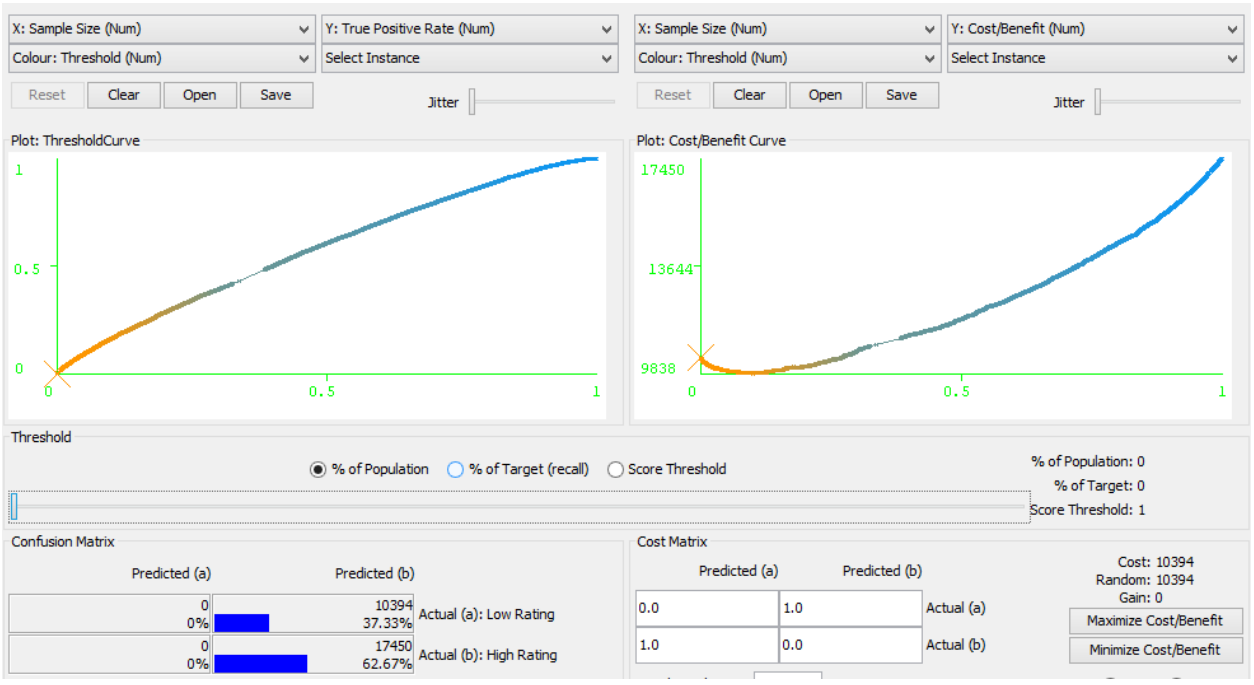
Classification Error:



Cost Benefit Curve: (For High Rating)



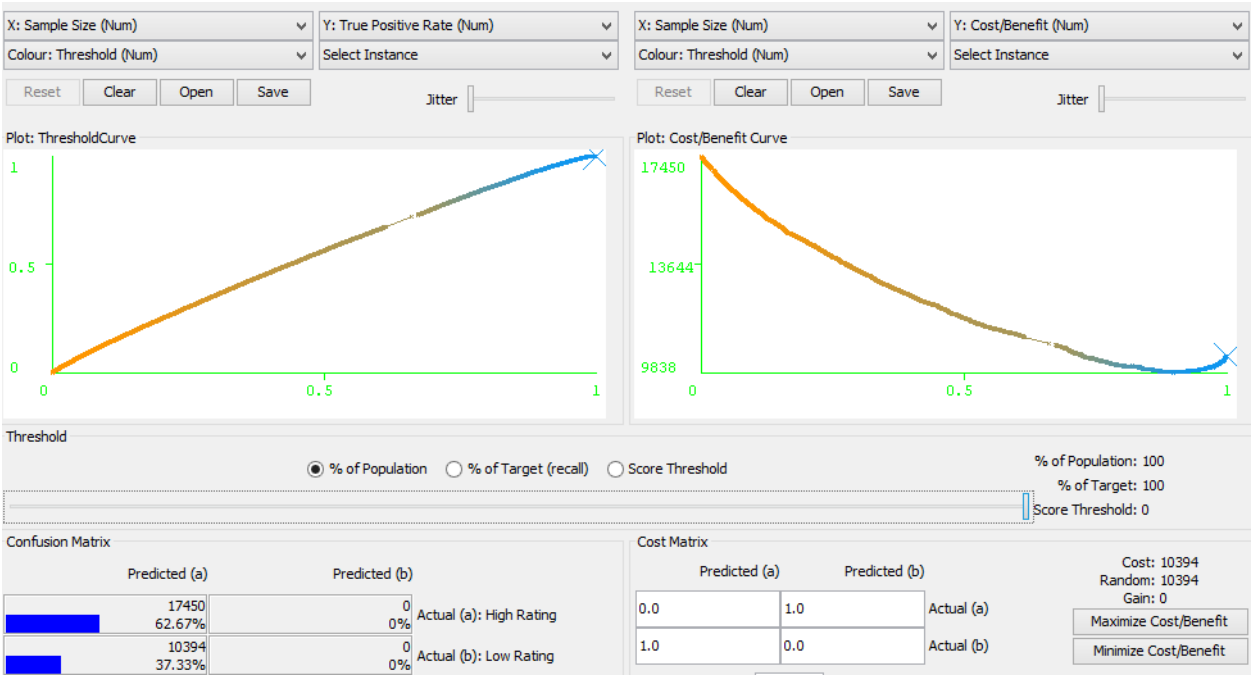
Cost Benefit Curve: (For Low Rating)



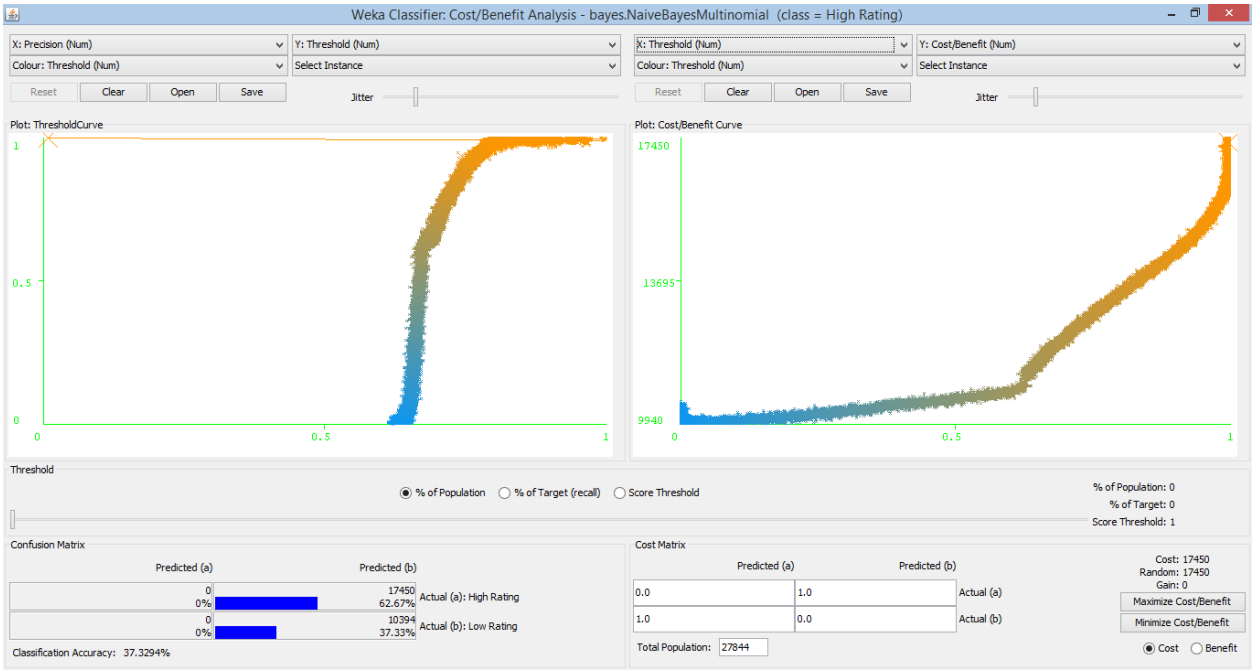
Task-2- Approach -2: (N-gram)

Multinomial Naïve Bayes

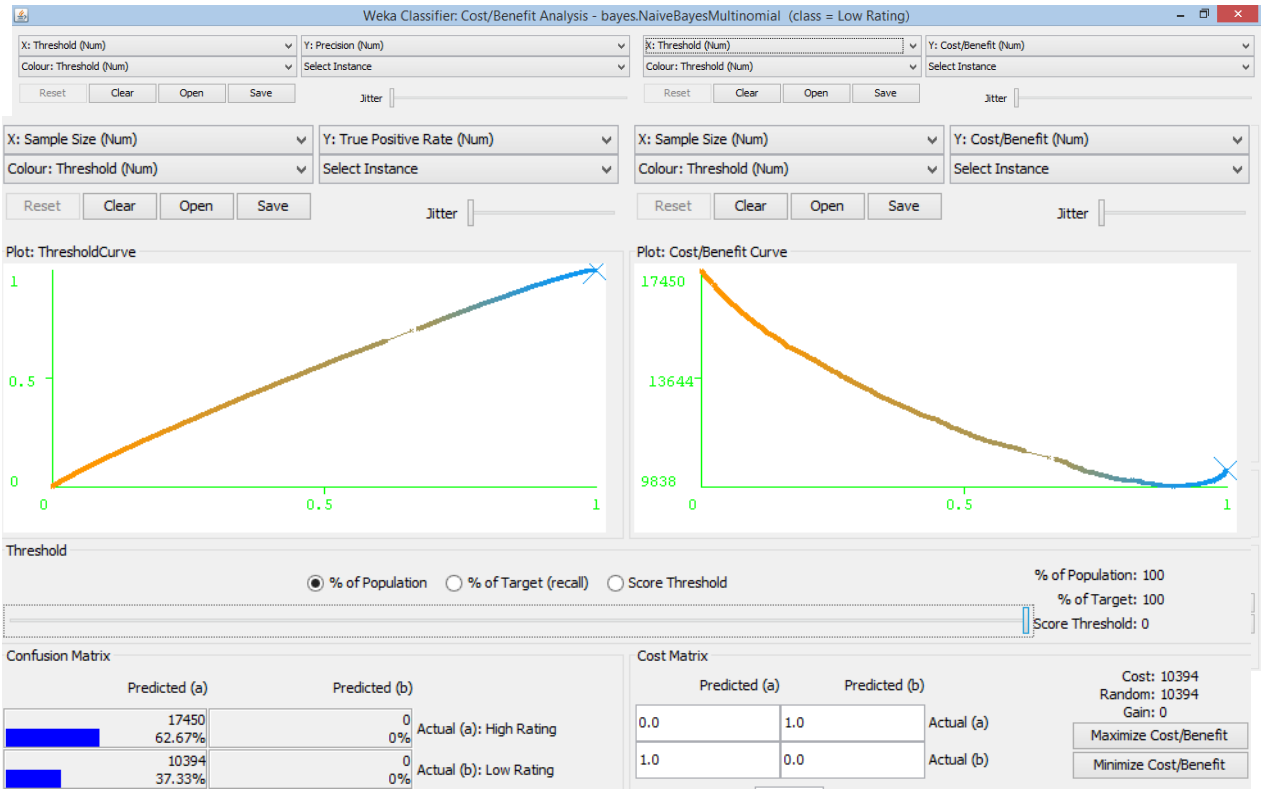
Classification Error:



Cost Benefit Curve: (For High Rating)

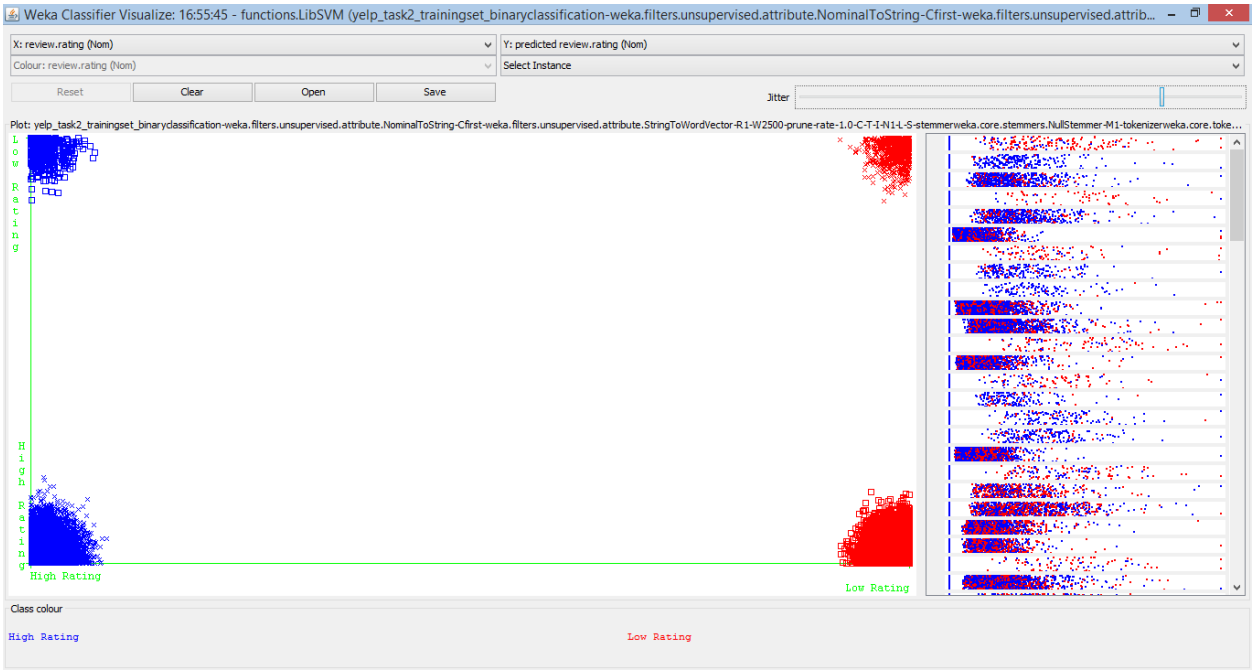


Cost Benefit Curve: (For Low Rating)

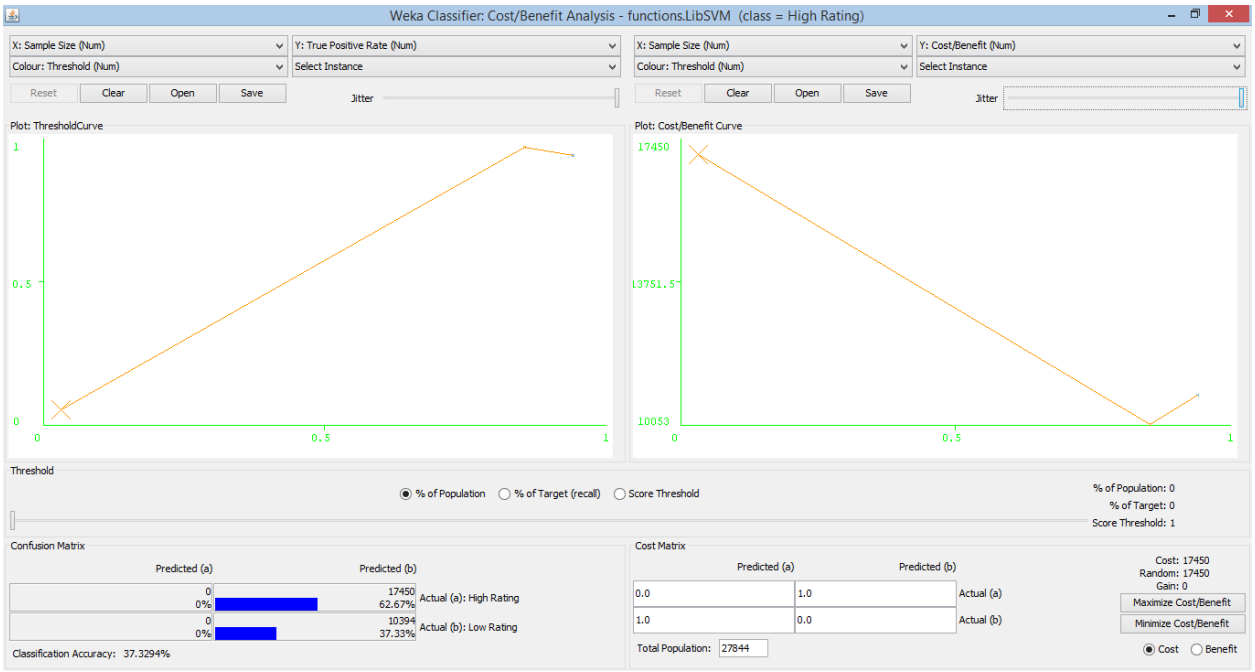


SVM's

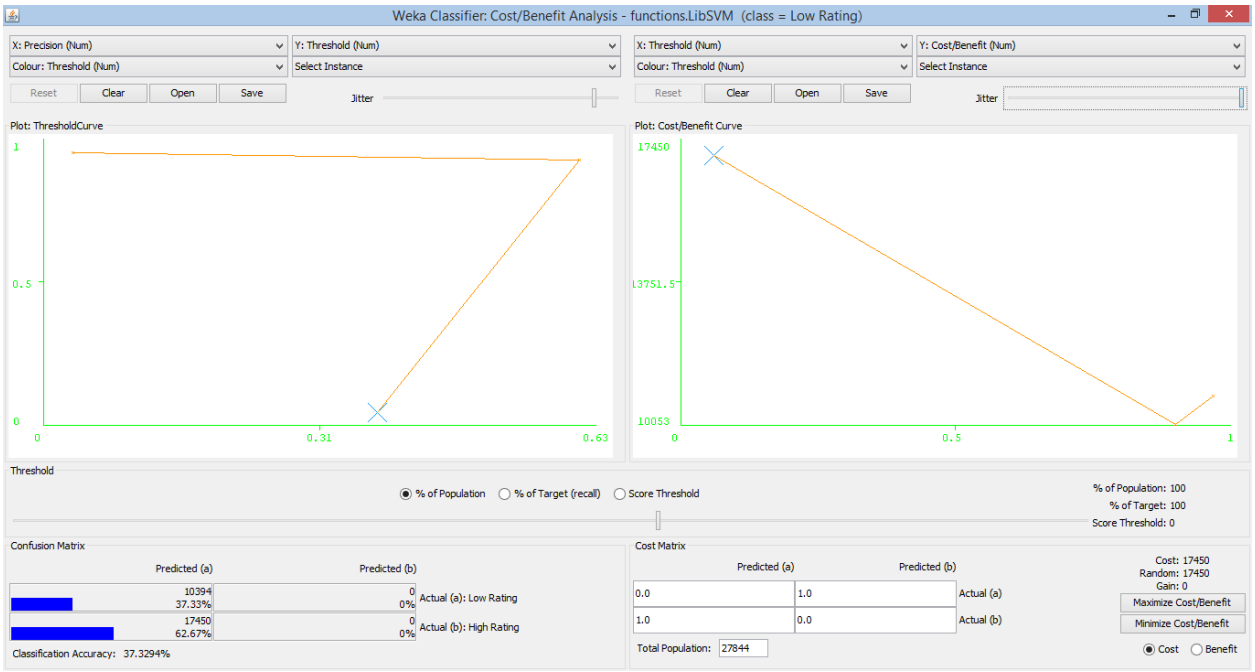
Classification Error:



Cost Benefit Curve: (For High Rating)



Cost Benefit Curve: (For Low Rating)



Analysis and Solution

Task-1:

Approach – 1:

Unigram's performed better than Bigram's and Trigram's across both classifiers as the number of relevant features is much higher. As the model gets more relevant information on Unigram features the accuracy and precision are high. Also an unigram can be thought of as a window placed over a text, such that only one word at a time is taken for modeling so the word independence is followed, even though the change in accuracy is not substantial it is still higher than bigram and trigram.

As far as the classifiers ae concerned, Multinomial Naïve Bayes Model performed better than Support Vector Machines. Number of instances plays some part in boosting accuracy when compared to SVM's, Naïve Bayes will perform better when there are smaller number of instances to validate.

Repeatedly resampling the data to form randomly partitioned training and validation sets, the average error on the training set became lower than on the validation set.

Evaluation Metrics shows that Precision and Recall are much higher for Unigram when compared to Bigrams and Trigrams.

Approach – 2:

In this approach as well, Unigram's performed similar to Bigram's across both classifiers as the number of relevant features is much higher. Trigram's performance couldn't be measured because of relative number of features (above 70000) which couldn't be projected into the Weka feature space.

As far as the classifiers ae concerned, Multinomial Naïve Bayes Model performed better than Support Vector Machines in this approach as well. When compared to Approach -1 performance without attribute selection is bit lesser. So, it can be concluded that performance with attribute selection works better.

Evaluation Metrics shows that Precision and Recall are much higher for Bigram's when compared to Trigrams and Unigrams.

Approach – 3:

In this approach Review information is added along with tip as well, Unigram's performed better than Bigram's and Trigram's as the number of relevant features is much higher. Only SVM's are used to classify the categories and it is evident that the performance is much higher when compared to only Tip information in Approach 1 and 2.

Evaluation Metrics shows that Precision and Recall are much higher for Unigram when compared to Bigrams and Trigrams.

Task-2:

Approach – 1:

Unigram's performed better than Bigram's whereas Trigram's performed better than the two tokenizers across Multinomial Naïve Bayes but it is less when compared to SVM's. As the number of relevant features is much higher. As the model gets more relevant information on Unigram features the accuracy and precision are high.

As far as the classifiers ae concerned, Multinomial Naïve Bayes Model performed less than Support Vector Machines.

The average mean root square error on the trigram and bigram training set became lower than on the unigram training set.

Approach – 2:

In this approach as well, when compared to Approach -1 performance with subset of N-grams (unigram, bigram and trigram) is bit lesser. So, it can be concluded that performance with N-grams separately worked better than subset of them in this case.

Conclusion

As Machine Learning approach is used for performing the Text Classification of “TIP” and “Review” information now when a new “TIP” comes in the category can be predicted to the level of accuracy the classifiers produced.

The same can be said about the Rating a particular ”Review” is going to award for the particular business.

Future scope lies in adding NLP methodologies such as POS tagging to improve the classifying accuracy of the predictive model.

Future Work

As Machine Learning approach supervised Learning is used for Text Classification. In Future Unsupervised Learning such as K-means Clustering can be used to study the performance.

References

- Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes Mulltinomial Naive Bayes for Text Categorization Revisited.
- Joachims, T: Text categorization with support vector machines: Leaning with many relevant features. In Proceedings of the Tenth European Conference on Machine Learning, Springer-Verlag (1998) 137-142
- T. Mitchell. Machine Learning. McGraw-Hill, 1997.
- Zhang, T., Oles, F.J.: Text categorization based on regularized linear classification methods. Information Retrieval 4 (2001) 5-31
- http://weka.wikispaces.com/Text+categorization+with+WEKA
- https://github.com/Yelp/dataset-examples/blob/master/json_to_csv_converter.py