CT-5103 Case Study for Data Analytics

Assignment 2

Classifying movie reviews: Crowd powered machine learning

Group Members:

1 Sai Krishna Lakshminarayanan (18230229)

2 Surya Balakrishnan Ramakrishnan (18231072)

Work Split Up

Student 1: Sai Krishna Lakshminarayanan (18231072)

- Task 4
- Task 5
- Task 6
- Half of Task 7
- Half of Task 8
- Task 10

Student 2: Surya Balakrishnan Ramakrishnan

- Task 1
- Task 2
- Task 3
- Half of Task 7
- Half of Task 8
- Task 9

Part 4 – Data description and result comparison

Introduction:

The aim of the assignment is to use the concept of using crowdsourcing data to obtain reviews for different movies. In this assignment we use 3 different approaches namely the gold standard approach, majority voting and finally the David and Skene method. We aim to compare the three models and decide which is the best model to compute reviews for a given movie. In the gold standard method we are given with a dataset which has a data of gold standard which is used to train the model and based on the training from the gold standard model we have the crowdsourced dataset for which the model tries to accurately classify the reviews into positive sentiment and negative sentiment in other words good or bad. We have used the decision tree classifier to accomplish the task. In the majority voting method, we consider all the reviews given to a movie and then the final target variable is positive if majority of the votes or sentiments are positive and the final target variable is negative if the majority of the votes for the movie is negative. The last and final method is the David and Skene's method. This is a weighted method where for each of the review we assign some weight which is the value of the significance. A review receives more weight if the review's sentiment matches the sentiment of the majority class and a particular review received less weight of the review's sentiment does not match the majority class.

Task 9

Characteristics of samples:

In the task 1, we have the gold.csv which contains a few movies and a set of reviews for the movies. From each review is a sentence from which a set of 1200 features have been extracted which represents a specific topic about the review. These features represent the sentiment of the reviewers for a particular movie. The id represents the movie, and topic columns represent the score received by the movie for a parameter. We have created a sample of 1000 rows which is saved into a csv file titled gold_sample.csv. The sampling methodology used here is simple random sampling.

Fig 1 Confusion Matrix

From the confusion matrix we can observe that the density of true positive and true negative is high as compared to the false positive and false negative. We can see that among the correctly predicted classes the decision tree model had a better accuracy predicting the true negative as compared to the true positives, which can be observed from the dark colour of the false negative section of the confusion matrix.

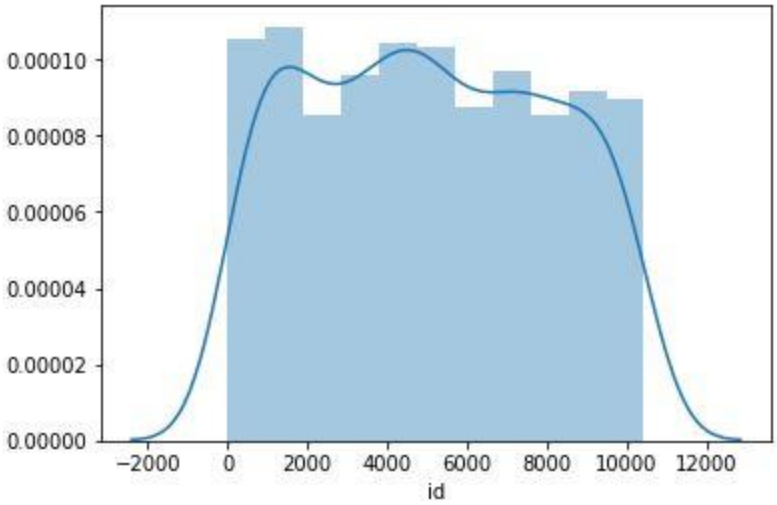


Fig 2 Distribution Plot

The distribution plot signifies what is the sentiment analysis score for each of the movie ids. We can see that there is no particular distribution followed by the distribution plot which signifies that the sampling of the data is truly done at random without any bias .

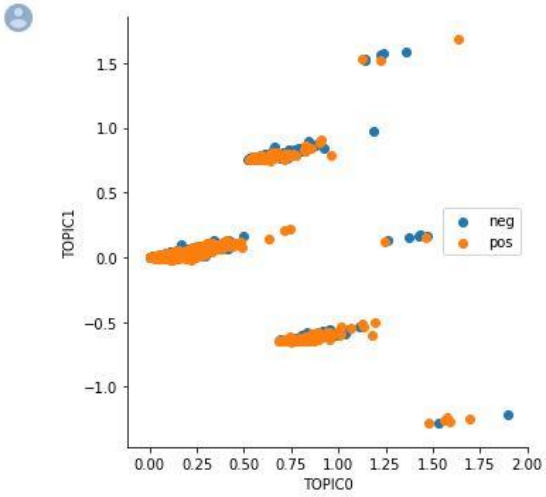


Fig 3 Scatter Plot.

From the scatter plot we can see that for each of the topics among the movie reviews which are regarded as an aspect of the review the correlation between the sentiment score by each individual with respect the final target class. Similar plots can be obtained for all the topics for the dataset which will helps to obtain some correlation between the sentiment score and the target variable or review being positive or negative.

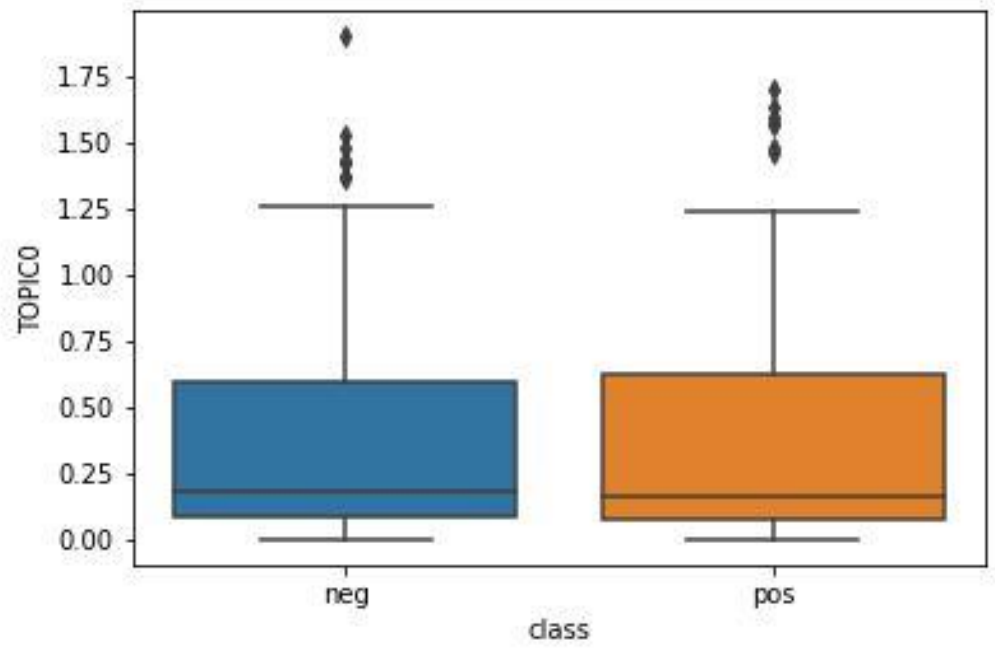


Fig 4 Box Plot

From the box plot we can see for each of the review topics what is the target class as for different values of sentiment score. Similarly the box plot also gives the amount of outliers in both the classes. From the box plot we can say that the positive target class has more outliers from which we can say that there is a good chance of the positive class having more fake reviews as compared to the negative target class.

In the task 4, we have the mturk dataset, which contains the reviews from actual reviewers for different movies. The dataset is merged with the generated dataset from task 1 based on the id column. Then the merged dataset is pre-processed to remove redundant columns. In the next step we take a sample from pre-processed dataset we use simple random sampling to get a sample of around 5000 and use aggregation techniques to group them by ids and adding the count of positive and negative sentiment after which the dataset is fed to the majority voting model.

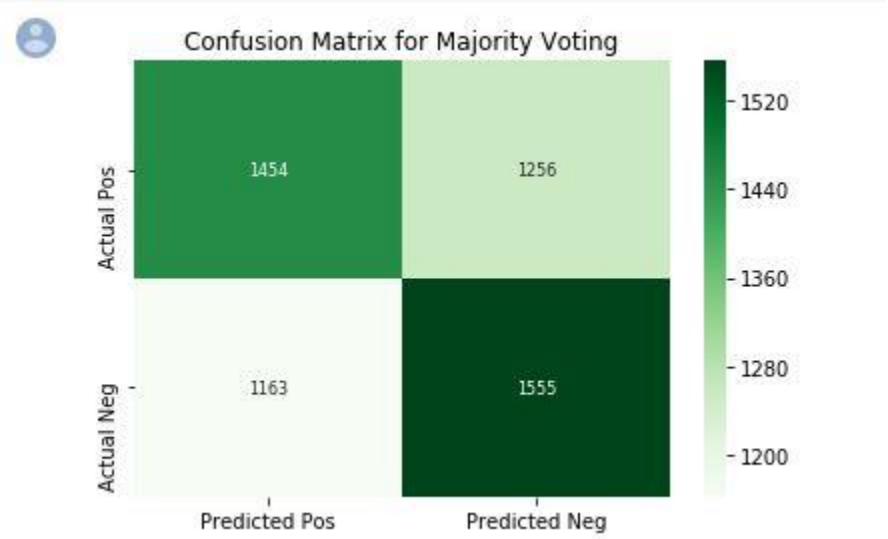


Fig 5 Confusion Matrix Majority Voting

From the confusion matrix we can see that as compared to the previous method in part 1 there are more true positive and true negative classified which suggests that the accuracy of the majority voting method is better as compared to simply feeding the dataset to the decision tree model. We can also observe that the density of true negative is greater as compared to the density of the true positive.

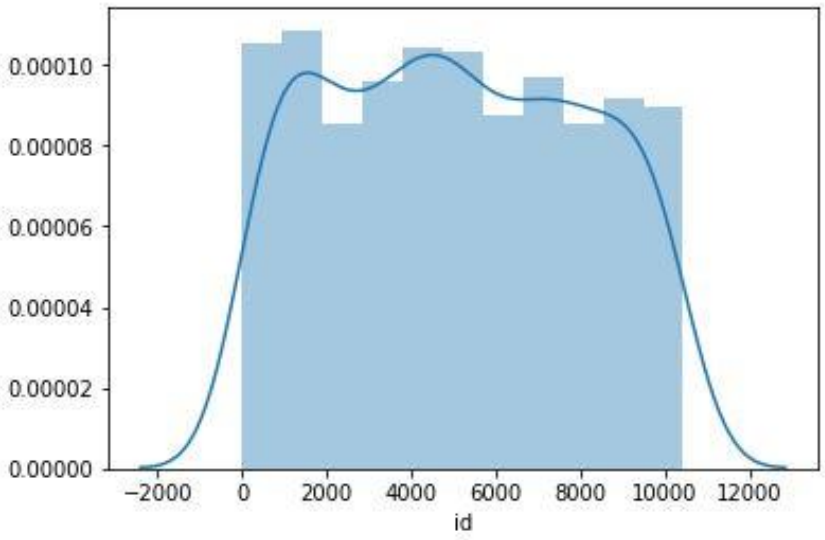


Fig 6 Distribution Plot Majority Vote

From the distribution plot we can see observe the sentiment score for each of the ids which is the different movies. Even in this case we can observe that the distribution is not normally

distributed which supports with the fact that the data was indeed randomly sampled without any bias in the sampling process.

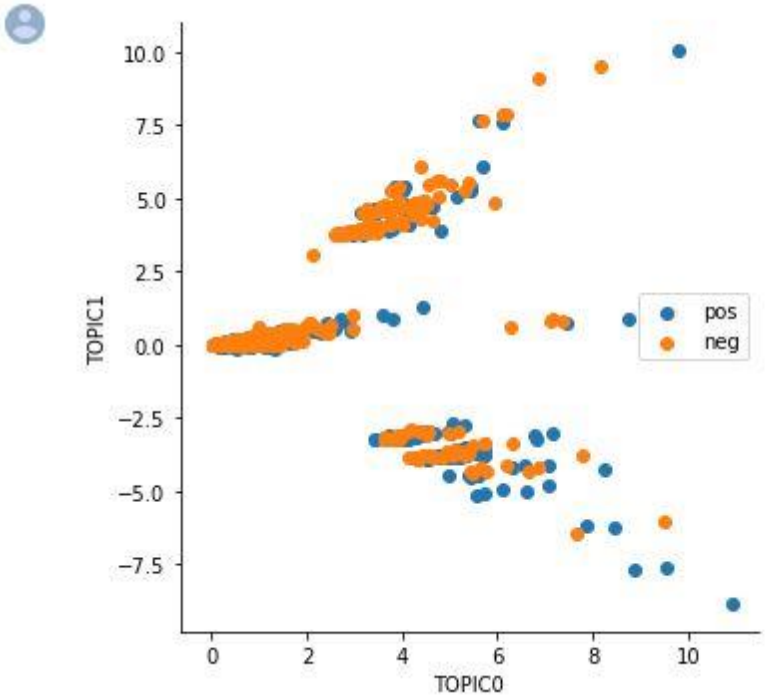


Fig 7 Scatter Plot Majority Vote.

From the scatter plot we can observe that there are more target classes which are classified as positive which suggests that the number of new true positives identified in this method is significant. This also suggests the fact the cluster of positive reviews increases as the sentiment score of the particular score increases. Similar plots can be obtained for all other topics of the movie review, which can help to obtain some correlation between sentiment score and the movie review being positive or negative.

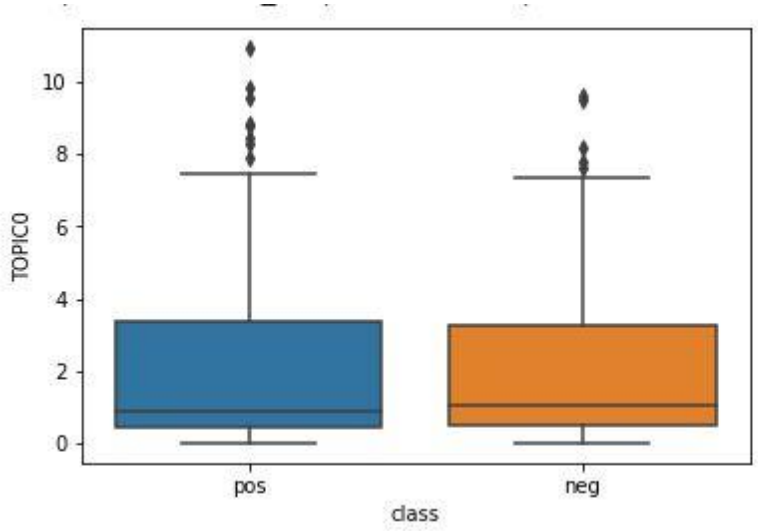


Fig 8 Box Plot Majority Vote

From the box plot we can see the plot between aggregated sentiment scores with respect to the movie review as positive and negative. From the box plot we can still observe that the number of outliers are more in case of the positive reviews from which we can say that the number of fake reviews in case of the positive class.

Using simple random sampling instead of selecting the first 1000 rows or rows of our choice ensures that the sample selected represents the unbiased proportion of the dataset. Simple random sampling ensures that each and every element that is being sampled has an equally likely chance of being picked as compared to every other element in the population. This ensures that we get a sample which is a true representation of the population.

Task 10

Description of the models

In part 1 we have trained the decision tree model based on the gold standard dataset. In the first step we take a simple random sample of 1000 rows from the gold.csv dataset. We have the x_train data frame which has the sentiment score for all the topics of the review, similarly we have the y_train data frame which contains the classification class which represents the review as positive or negative for a movie by a user which is the target variable. In the next step we assign and initialise the decision tree classifier and then we feed the x_train and y_train variables to the decision tree for fitting the classifier model. Then we feed in the test dataset to the decision tree classifier model. The decision tree model then predicts the value of y based

on the fed values. Then we obtain the confusion matrix, accuracy, precision and recall values for the decision tree model. We observe that the the model has an accuracy of 53% and considering the weighted average precision recall method we have a score of 53%. We also observe that we have a positive recall score of 55% and a negative recall score of 51%. From the confusion matrix we cans see that we have 25% true positive values, 24% false positive values, 23% true negative values, and 28% of false negative values.

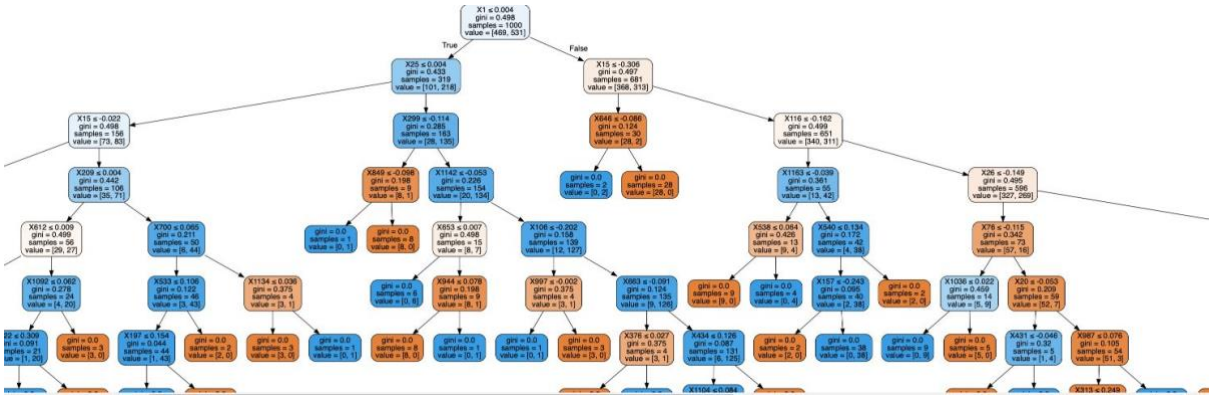


Fig 9 Decision Tree Gold Standard

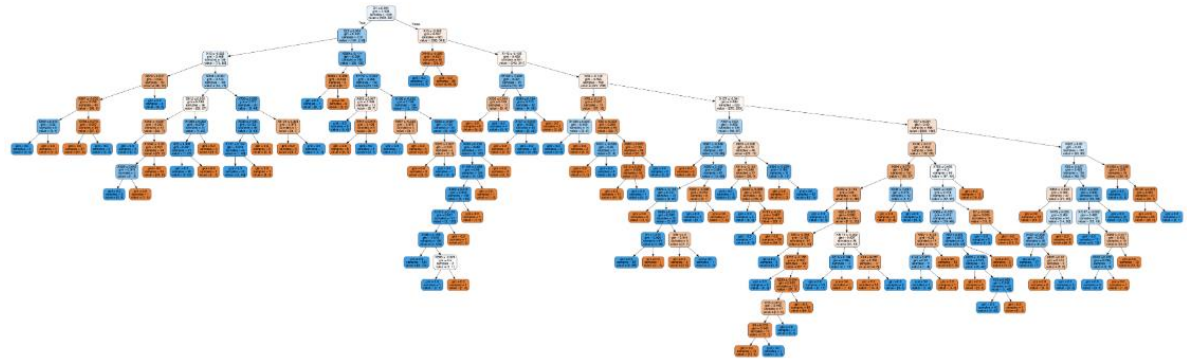
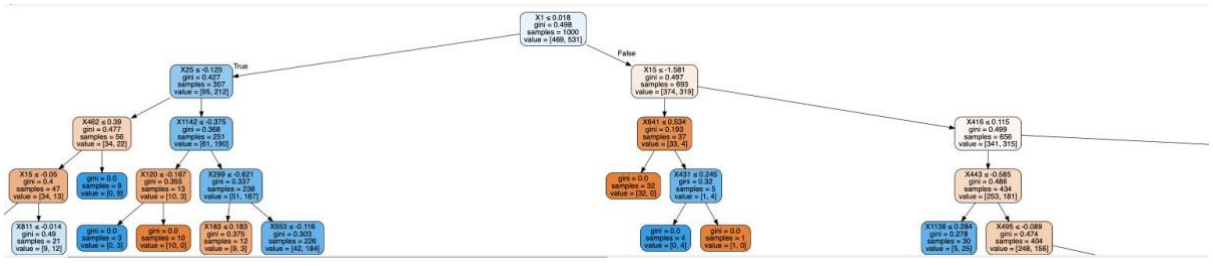


Fig 10 Decision Tree Gold Standard

This is the decision tree model which was obtained as the output in part 1. All the classes were classified based on the label values which were classified for different sentiment scores.

In part 2 we have trained the decision tree model using the crowdsourced data using the majority voting methodology. We create a positive column which will have its value as 1 the class is pos, similarly we create another column the negative column which will have its value



1 if the class is neg. For each of the columns positive and negative we calculate the count, which signifies the number of positive and negative reviews for a particular movie for the purpose of grouping. We use aggregation methodology and simple random sampling to obtain a sample which will be fed to the decision tree classifier. Similar to the task 3 we split the dataset into x_train and y_train, where x_train has the sentiment score for all the topics of the review and y_train will contain the target variable. Then we assign and initialise the decision tree model and feed it with an input and the test dataset and predicting the target variable. In the majority voting method the target variable is decided based on the majority votes. In other words if the number of positive reviews is greater than the negative reviews then the target variable is set to positive, else the target variable is set to negative. In the mterk dataset among the total of 5511 annotators there are 186 unique annotators among them there are a few fake ones which are names as fake0, fake1, and so on. We observe that the model has an accuracy of 55% and considering the weighted average precision recall method we have a score of 55%. We also observe that we have a positive recall score of 57% and a negative recall score of 53%. From the confusion matrix we cans see that we have 32% true positive values, 27% false positive values, 25% true negative values, and 34% of false negative values.

Fig 11 Decision Tree Majority Voting

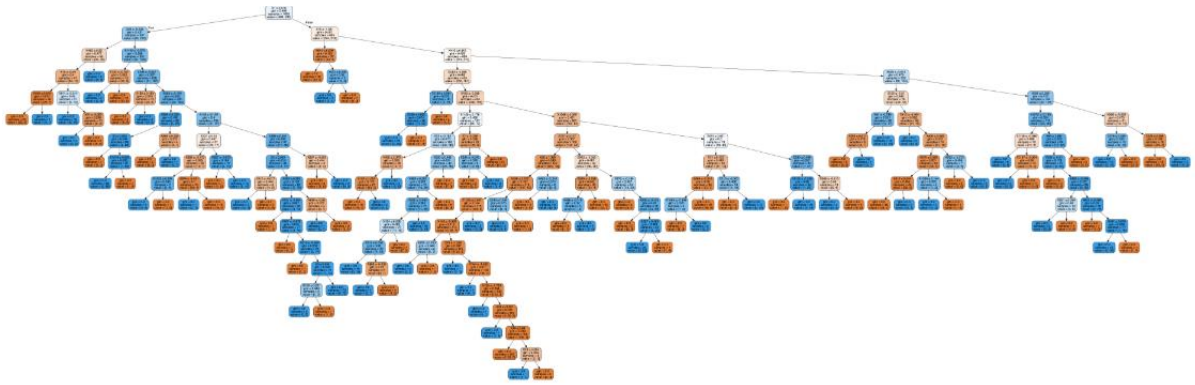


Fig 12 Decision Tree Majority Voting

This is the decision tree model which was obtained as the output in part 2. All the classes were classified based on the label values which were classified for different sentiment scores.

In part 3 we train the crowdsourced data using the David & Skene method. The David & Skene method has an approach which is weighted. The individual terms are assigned a weight which is a measure of significance. More the weight of an element more significant the term is. Generally in a movie review system there are people who tend to repeatedly under different

names or ids give positive or negative reviews. The David & Skene method is implemented over the majority voting method, where the entities whose target variable matches the majority class has more weights as compared to the entities which have a target variable different to the majority class.

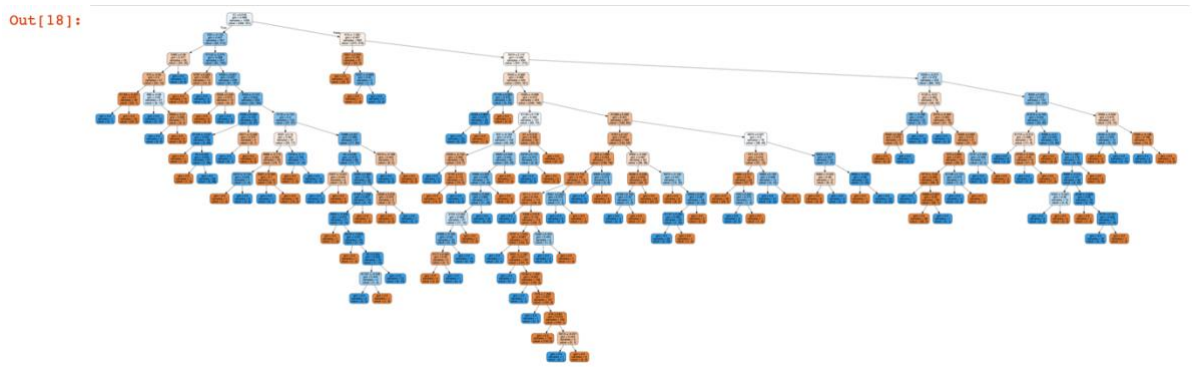


Fig 13 Decision Tree David and Skene

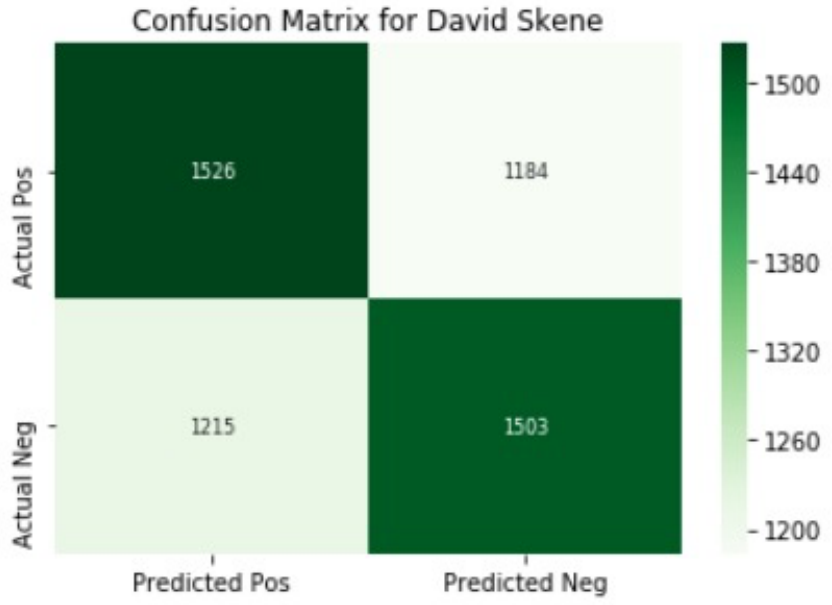


Fig 14 Confusion matrix David and Skene

In the David and Skene method we first take the dataset which was used for the previous case of majority voting. Then based of what the vote of the majority class is if for a particular

movie's topic if the majority class is positive and a particular reviewer has also given a positive review then we make sure that the review counts by increasing the significance. Similarly if a particular review is not matching with the majority class then we decrease the impact of the review. In this case we have considered weighted sum method. Then once the weighted sum is calculated then the target class is calculated based on the weighted sum. From the confusion matrix we can observe that the number of true positive and true negative is the highest among all the three models. We observe that the model has an accuracy of 56% and considering the weighted average precision recall method we have a score of 56%. We also observe that we have a positive recall score of 55% and a negative recall score of 56. From the overall results we can say that the David and Skene's method produces the best accuracy score.

References:

1 https://stackoverflow.com/questions/51569300/installing-graphviz-in-windows-10-with-anaconda3-and-using-it-through-a-jupyter?rq=1

2. https://stackoverflow.com/questions/34861086/replacing-null-values-in-a-pandas-dataframe-using-applymap

3. https://stackoverflow.com/questions/45416684/python-pandas-replace-multiple-columns-zero-to-nan

4. https://stackoverflow.com/questions/19482970/get-list-from-pandas-dataframe-column-headers

5. Lecture Slides and references mentioned in the lecture slides.