Abdullah Alhassan

Professor Shawn Bowers

CPSC 310: Data Mining

15 December 2016

**Airline Sentiment from Twitter Feeds**

**Introduction**

A 14640 twitter feeds were collected during February 2015. The feeds contained an airline mention, specifically these airlines: United, Virgin America, American, Delta, US Airways, and Southwest. Those tweets, then, were rated into three categories regarding a feedback: positive, neutral, and negative. Originally, the dataset contains 13 attributes shown in Table 1, then data cleaning has been performed to on the dataset to make it useful. A major cleaning is removing the key attributes such as tweet id, and other attribute unnecessary attribute that does not help classify an airline sentiment such as name, tweet\_coor, and user\_timezone. The only attributes left that help generate useful classifiers are airline, retweet\_count, tweet\_created, and airline\_sentiment. The tweet\_created attribute was given in date-time format e.g “YYYY/MM/DD HH:MM”, which was converted to a day of week name that will help categorizes airline performance according to the day of week. In this project there were five types of classifiers ensemble, decision tree, k-nearest neighbor, linear regression, and naive bays.

|  |  |
| --- | --- |
| **Attribute** | **Explanation** |
| tweet \_id | Is the unique key for every tweet |
| negativereason | Negative comments of the flight |
| airline | The name of the airline |
| airline\_sentiment\_gold | Is the result after testing the dataset on a classifier |
| name | The name of person tweeting |
| negativereason\_gold | Is the result after testing the dataset on a classifier |
| retweet\_count | Number of retweets |
| text | The actual tweet |
| tweet\_coord | The coordinates of the tweet |
| tweet\_created | The date of when the tweet was created |
| tweet\_location | The location of the tweet |
| user\_timezone | The time zone of the user |
| airline\_sentiment | The class label: positive, neutral, and negative |

Table 1: Dataset Attributes

**Data Analysis**

The dataset after cleaning contains four attributes: airline, retweet\_count, tweet\_created, and airline\_sentiment. The class attribute is the airline\_sentiment, which will be classified using the number of retweet, day of tweet, and airline name. The total number of instances is 14640 and it does not contain any missing value.

The class attribute, airline\_sentiment, has three different values that determine the rating of the airline, those values are “positive”, “neutral”, and “negative” in order of best to worst. Hence, the class attribute is an ordinal. The retweet\_count is an indicator of the number of people who agreed with the tweet review of the airline, which means the higher the retweet the more accurate the sentiment of the airline. Therefore, the retweet\_count attribute is a ratio with zero meaning that no retweet has happened. The day of the tweet, which was derived from the original tweet\_created value is considered to be ordinal value, because it is a set of the week days. Finally, the airline attribute which contains a set of six airline companies that are being rated from tweet feeds.

The attributes can be more understood with data visualizations. First, considering the number of instances in the dataset a frequency diagram can help illustrates the tendency of the data. In Figure 1 shows a frequency diagram for airline names. Accordingly, Virgin America airline has got the lowest feeds, as a result its sentiment review is the least accurate. On the other hand, United airline has the maximum twitter feeds, which indicates that it has the most accurate sentiment results.



Figure : Frequency diagram for airlines

Figure 2 shows frequency diagram for the day of tweet, and the highest tweet days are Sunday and Monday, and the lowest days are Wednesday and Thursday. The week day frequency is helpful to understand what day of the week has the most effect on the classification process, and according to Figure 2 Sundays and Mondays are dominant.



Figure : Frequency diagram for day of tweet

The retweet count is an indicator to how accurate the review is. However, Figure 3 indicates that almost all the tweets were never retweeted, and under 1000 tweet were tweeted once. For this reason, the sentiments are considered individual opinion in this dataset.



Figure : Frequency diagram for retweet\_count

The public opinion, from twitter, of the airlines can be shown in Figure 4 where a frequency diagram for the airline sentiment is shown. According to Figure 4, two thirds of the airline sentiments are negative, which is a problem for the airline corporations to study and improve airline accommodations.



Figure : Frequency diagram for airline sentiment

For further analyses, a multi-frequency diagram is shown in Figure 5 that illustrates the frequency for the week days for each airline. All of the airlines have feeds for every day of the week, except for American airlines that only has feed for Sunday, Monday, and Tuesday, which can be a problem for the accuracy of the reviews. However, American airlines has the most review count for the specific days it has review on, which can even out the lack of review for other days. The United airlines shows the best review distribution for the week days, since almost every day has a sufficient amount of review for the sentiment to be accurate.



Figure : Multi-Frequency diagram for days and airlines

The last data visualization analysis is shown in Figure 6 where it demonstrates the multi-frequency diagram for airlines and their sentiments. The first thing to notice in Figure 6 is that all of the airlines have the negative sentiment as the highest count, which means that people are not happy about the airlines treatments. Therefore, to make sense of the dataset, the ratio of the negative sentiment to the total count of the sentiments for each airline has to be calculated, the ratio formula is given in Figure 7. Hence, the airline with the highest ratio is considered the airline with the best review. Table 2 shows every airline and their ratios. According to Table 2, Virgin America has the best rate, unlike US Airways that has the worst rate.



Figure : Multi-Frequency diagram for airlines and their sentiments

Figure : Ratio Formula

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Airline | Positive count | Neutral count | Negative count | Total | Ratio |
| Virgin America | 152 | 171 | 181 | 504 | 64.09% |
| Delta | 544 | 723 | 955 | 2222 | 57.02% |
| Southwest | 570 | 664 | 1186 | 2420 | 50.99% |
| United | 492 | 697 | 2633 | 3822 | 31.11% |
| American | 336 | 463 | 1960 | 2759 | 28.96% |
| US Airways | 269 | 381 | 2263 | 2913 | 22.31% |

Table : Detailed sentiment counts and the ratio of each airline

**Classification Results**

As mentioned earlier in the introduction, there were five types of classifiers used for the dataset ensemble, decision tree, k-nearest neighbor, linear regression, and naive bays. In this section, more detailed will be discussed for each classifier.

The first classifier used was the ensemble with F set to 2, N set to 50, and M set to 37. The ensemble used was of homogenies type that uses N decision trees, in this case 50 different decision trees. After constructing 50 decision trees 37 were chosen using stratification technique. The way that ensemble choses a class for an instance is by taking the majority voting for all 37 decision trees. In case, there was no leaves in a decision tree, the ensemble algorithm assigns negative class to the instance, because most of the data provided are negative. Figure 7 shows the predictive accuracy of the ensemble classifier. The accuracy of the ensemble is above 50%, which is a considered a decent classifier. To further the analysis of the ensemble, Figure 9 shows the confusion matrix for this classifier. According the confusion matrix, the ensemble classifier is bad at classifying positive instances, which is expected because the positive sentiment is the least frequent of all sentiments.

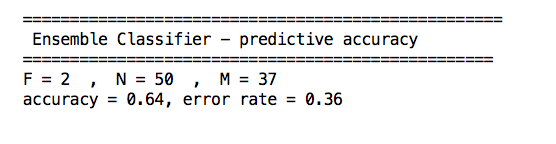


Figure 7: Ensemble classifier accuracy result

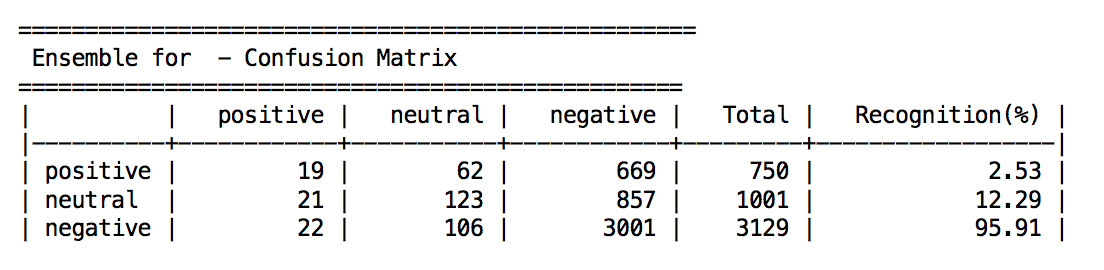


Figure : Ensemble confusion matrix

As oppose to use multiple decision trees, the next classifier is one decision tree. The accuracy result of the decision is equal to the accuracy of the ensemble, as shown in Figure 9. It is expected for the decision tree to have equal or close predictive accuracy as the ensemble, since the ensemble uses multiple decision trees of the same algorithm. However, Figure 10 shows the confusion matrix of the decision tree, which is worse than the ensemble. The confusion matrix of the decision tree is bad at guessing both positive and negative sentiments. Even though the predictive accuracy of both the ensemble and the decision tree are equal, the ensemble classifier is slightly better classifier, because of the confusion matrix results.

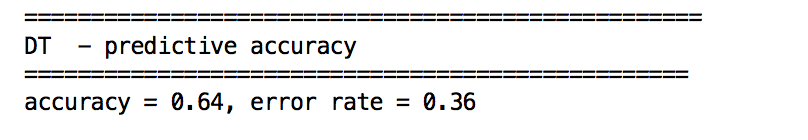


Figure : Decision tree predictive accuracy

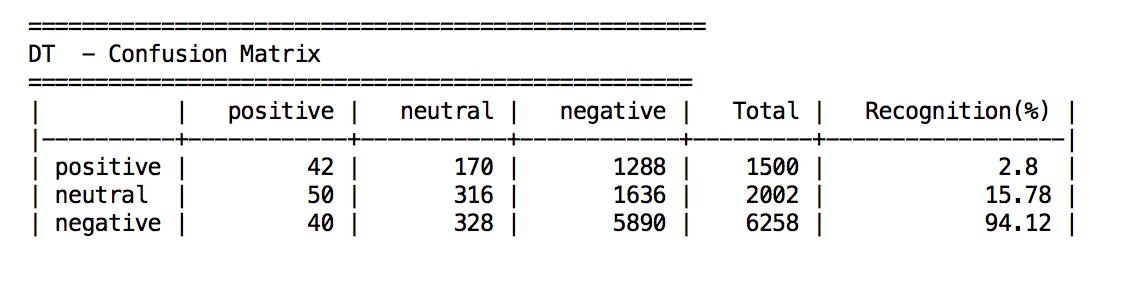
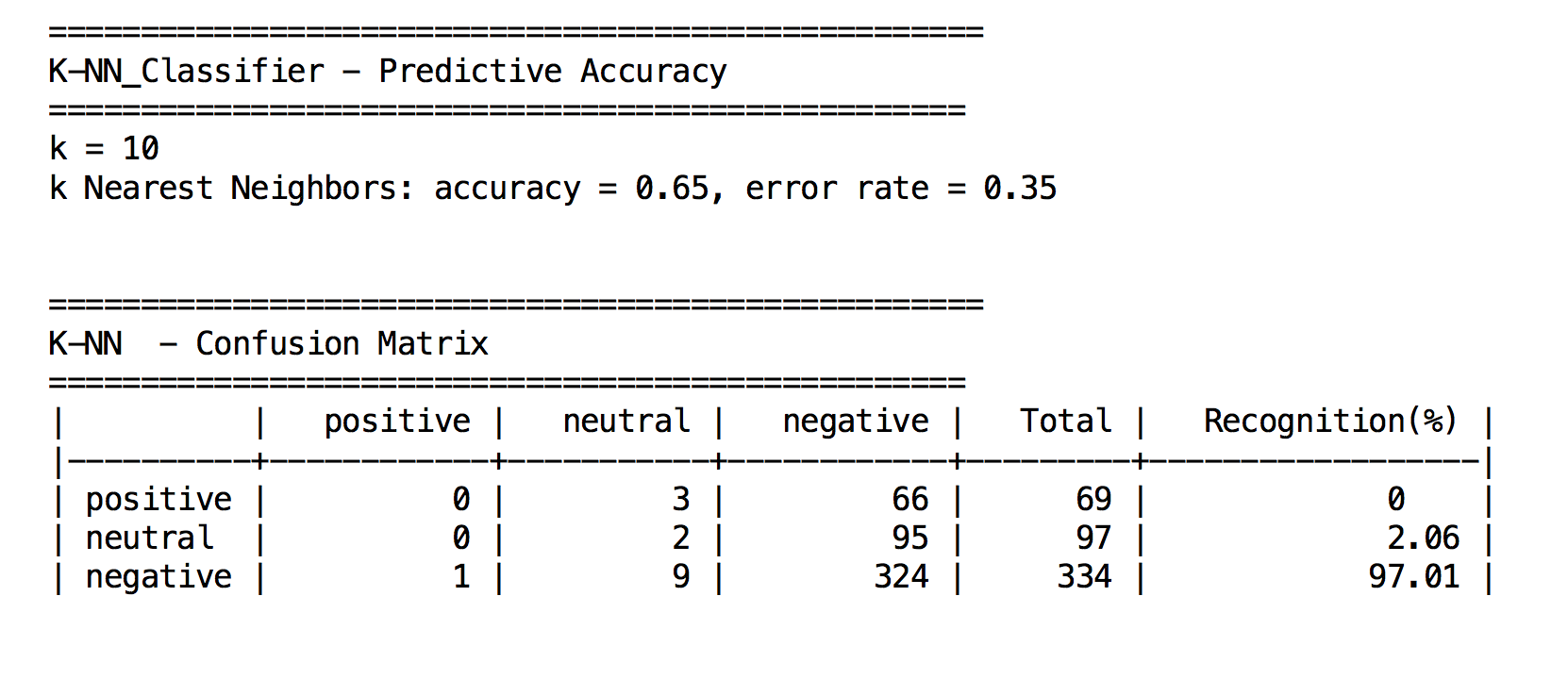


Figure : Decision tree confusion matrix

The next classifier used was k-nearest neighbor, and its predictive accuracy is shown in Figure 11. Although the accuracy is above 50%, K-NN classifier was tested only on 500 instances due to the amount of the dataset. the K-NN algorithm used could not handle this amount of the dataset, for it took a considerable amount of time to process all of the data, which is the reason only 500 instances were tested.

Figure : K\_NN predictive accuracy and confusion matrix



The fastest classifier used for this dataset is the linear regression, after it processed the dataset the algorithm only chose negative sentiment because of its dominance over the others. Hence, for a classifier that only selected negative sentiment it made a 62% accuracy as shown in Figure 12, which also shows the confusion matrix and how it only guesses negative sentiment.

The last classifier used was the one that depends on probabilities, Naive Bays. As all of the other classifiers, Naive Bays has predictive accuracy above 50% as shown in Figure 13. Although just like K-NN due to the dataset size, naïve bays was only tested on 500 instances.

Figure : Linear regression predictive accuracy and confusion matrix

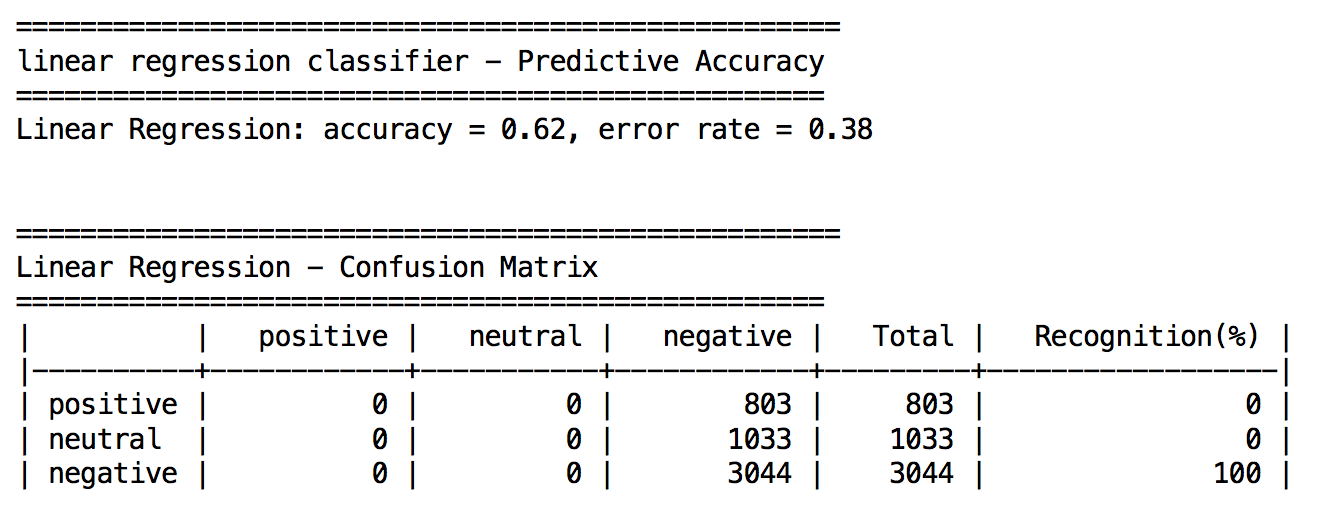
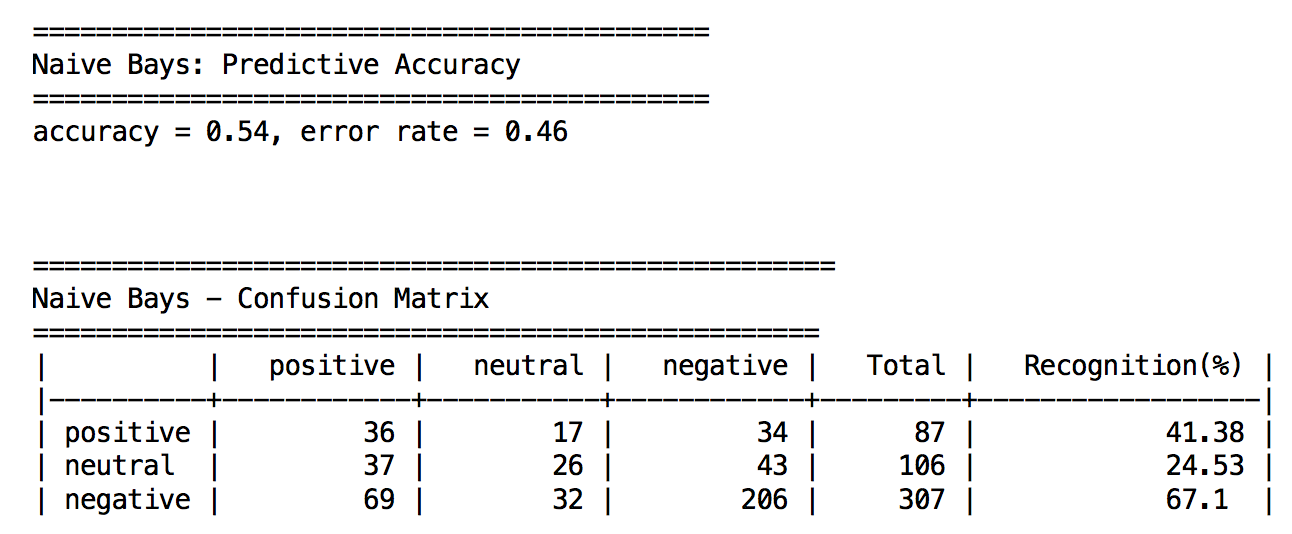


Figure : Naive Bays predictive accuracy and confusion matrix



**Conclusion**

In brief, the dataset is a twitter feed that mentions an airline, then the tweet itself is evaluated to one of the following sentiments: positive, neutral, or negative. The sentiment is the attribute to predict in this project. The different classifiers used show approximately close predictive accuracy number, which indicates that all the classifiers used have a correct algorithm of prediction. One of the challenges faced in this project is to select the right attributes to classify that make the prediction logical. In this project, given the day, the airline, and the retweet count then the classifiers will predict a sentiment for the airline chosen. The location of the tweet is a valid attribute to include, but the values given in the dataset for the location are not convenient to use they are either invalid or unavailable.