Sentiment Analysis

Using SAS Sentiment Analysis Studio and R Studio

By Sazee S. (Maninderpreet Singh Puri)

Table of Contents

1. Introduction- Case Study B

The Case Study B is a sentiment analysis assignment on the dataset *'apple_review_new.csv'* which consists of 2 variables with 143 observations. The goal is to develop a dictionary-based sentiment analytics engine based on the R library *'syuzhet'* and *'tidytext'* to analyse the different emotions from Apple review tweets. We are also developing a machine learning-based model using the R libraries *'tm'* and *'e1071'* as well as evaluating the predictive accuracies of SVM classifier. Also, we intend to develop a statistical model using SAS Sentiment Analysis studio and evaluate the accuracies.

2. Task 1

After combining the two variables ('*positive'* and '*negative'*) into one single column I plotted the sentiment change with respect of narrative time. We can see that positive review are plotting above the x-axis and negative reviews are plotted below the x-axis.

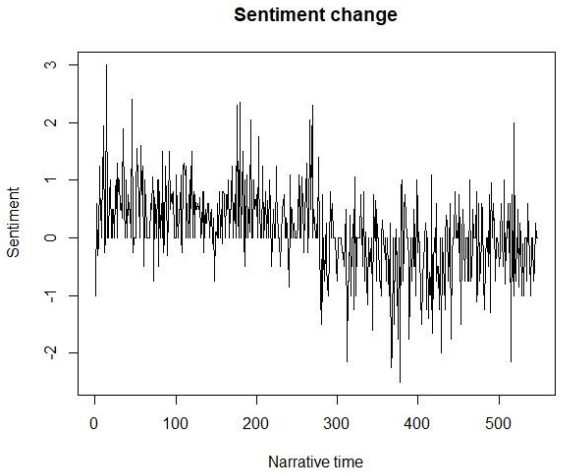

Figure 1 Correlation matrix of Entertainment domain.

After applying the dictionary-based sentiment analysis engine based on R library *'syuzhet'* and *'tidytext'*, I got a plot of sentiments (*anger, anticipation, disgust, fear, joy, sadness, surprise, and trust*).

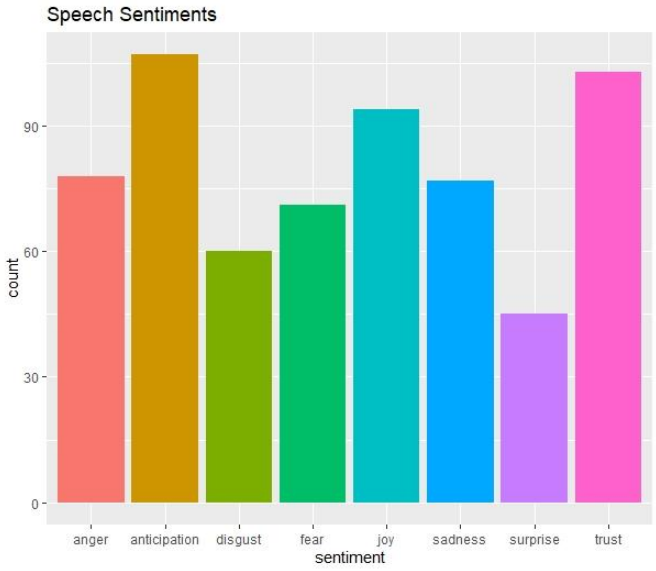

Figure 2 Bar plot showing the distribution of sentiments in review words by customers.

The plot shows that most of the words are related to anticipation, trust and joy which shows that most of customers are happy and anticipate new products from Apple. However, there are other customer who showed disgust, fear, and sadness in their reviews. Only a few customers expressed surprise sentiment in their reviews.

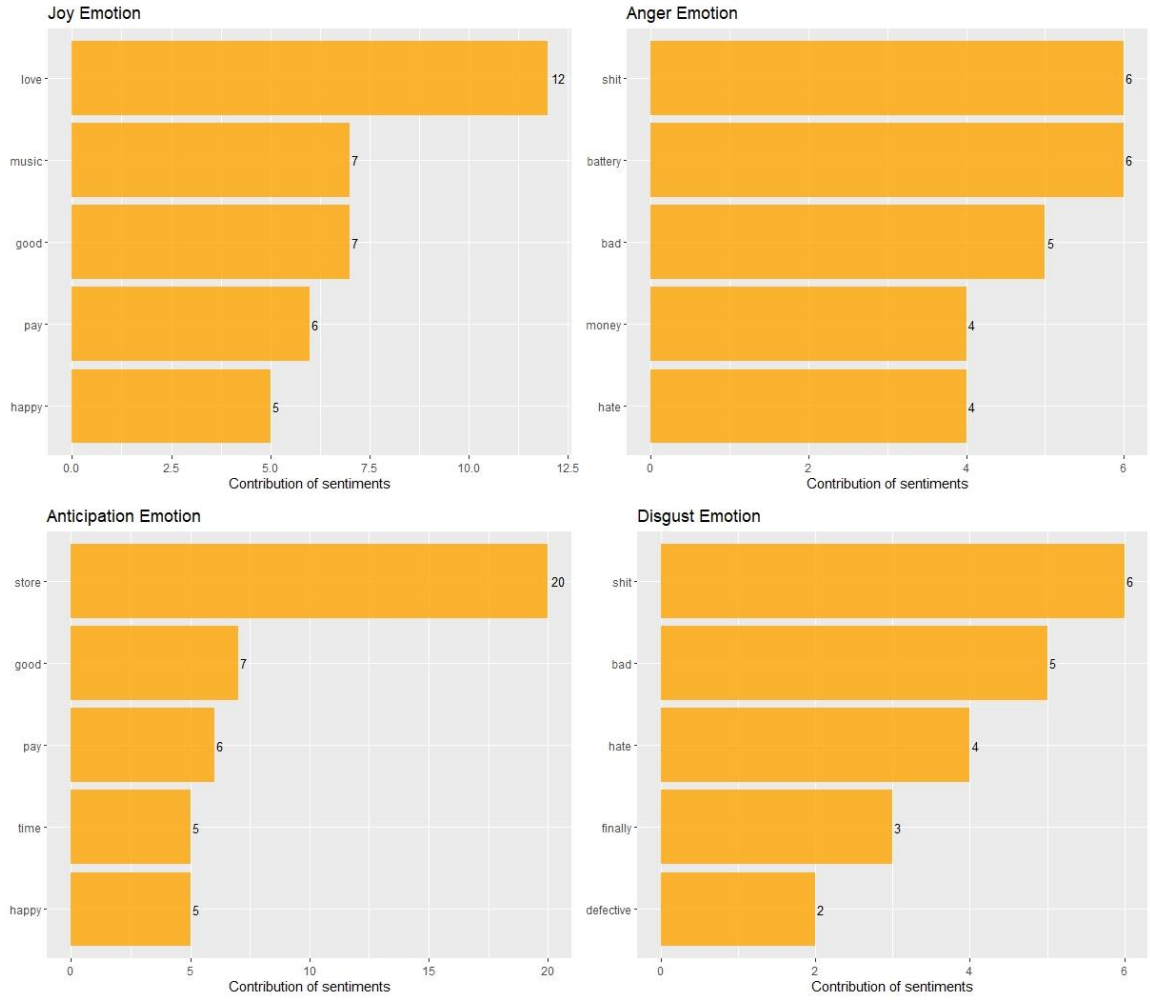

Figure 3 Bar plot showing the sentiments the top 5 most frequent words in each of the eight emotions.

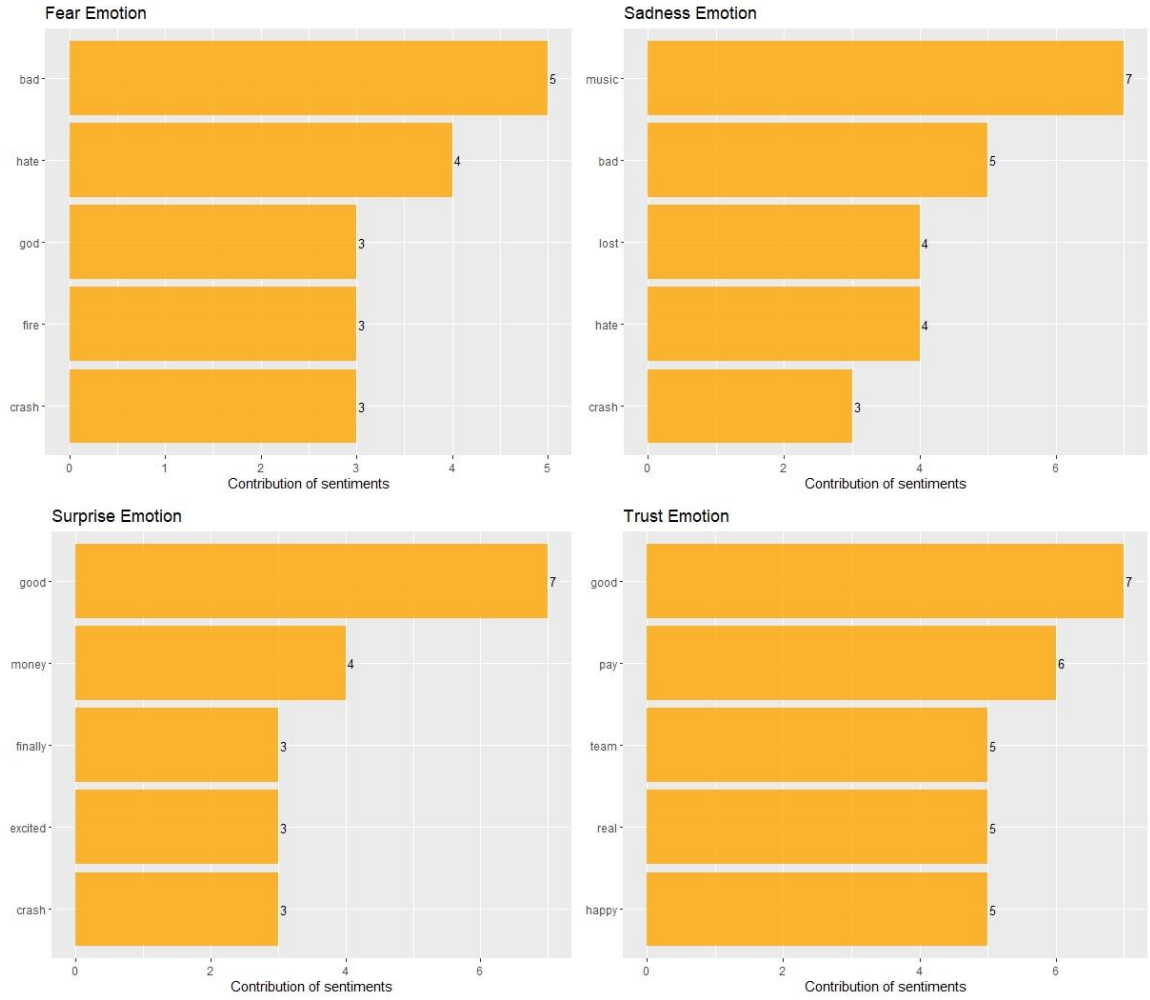

Figure 4 Bar plot showing the sentiments the top 5 most frequent words in each of the eight emotions.

We can see that the analysis shows that the top frequent words in each of eight emotions are relevant to their respective categories like in '*joy'* emotion the top word is *'love'* followed by '*music'* which both show '*happy/joy'* emotions. These plots also tell some words are used more in review than other in the same category and these words have a huge difference in their frequencies. For instance, in the '*anticipation'* category, we can see that '*store'* is occurring twenty times, while the word *'happy'* occurred five times.

3. Task 2

In this task, I applied State Vector Machine (SVM) method to predict sentiments (positive and negative). After splitting the data set into training (80 %) and testing (20 %), I applied the SVM classifier to build to model. I tested the model with testing dataset and got 83.3% accuracy.

```
Confusion Matrix and Statistics

          Reference
Prediction negative positive
  negative       18        2
  positive        6       22

               Accuracy : 0.8333
                 95% CI : (0.6978, 0.9252)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 1.653e-06

                  Kappa : 0.6667

 Mcnemar's Test P-Value : 0.2888

            Sensitivity : 0.7500
            Specificity : 0.9167
         Pos Pred Value : 0.9000
         Neg Pred Value : 0.7857
             Prevalence : 0.5000
         Detection Rate : 0.3750
   Detection Prevalence : 0.4167
      Balanced Accuracy : 0.8333

       'Positive' Class : negative
```

Figure 5 SVM Classifier prediction results.

The results also show the model predicted six misclassifications for the negative sentiment and two for the positive sentiment. According to the R documentation ("Package 'caret," 2021) the precision score is 0.9 and recall score is 0.75. We can say by the scores that the model is performing well in predicting the sentiments in the reviews.

4. Task 3

In this task I used SAS Sentiment Analysis studio to predict the sentiments (positive or negative) in the apple reviews dataset. There are many parameter settings in SAS Sentiment Analysis studio. I tried many combinations to find the best precision.

Table 1 Table show the training results of models with different parameter setting.

Model(s)	Simple/advance	Best model (check/uncheck)	Training/ testing ratio	Probability threshold	Text normalization model	Feature ranking algorithm	Overall Precision	Positive Precision	Negative Precision
Model 1	Simple	Uncheck	80/20	Default (0.50)	Default (Smoothed Relative Frequency)	Risk Ratio	91.67%	95.83%	87.50%
Model 2	Simple	Checked	80/20	Default (0.50)	Smoothed Relative Frequency	Risk Ratio	91.67%	95.83%	87.50%
Model 3	Advanced	-	70/30	Default (0.50)	Smoothed Relative Frequency	Risk Ratio	86.11%	86.11%	86.11%
Model 4	Advanced	-	90/10	Default (0.50)	Smoothed Relative Frequency	No Feature Ranking	91.67%	100.00%	83.33%
Model 5	Advanced	-	90/10	(0.30)	Smoothed Relative Frequency	No Feature Ranking	91.67%	100.00%	83.33%
Model 6	Advanced	-	90/10	(0.80)	Smoothed Relative Frequency	No Feature Ranking	91.67%	100.00%	83.33%
Model 7	Advanced	-	80/20	Default (0.50)	Okapi BM25	Risk Ratio	91.67%	95.83%	87.50%

The first phase is the training phase. As we can see in the table, the model one performs the best out of all the models we considered for our study. We can see that the overall precision in model one is the highest (91.67%), with the positive correlation (95.83%) and negative correlation (87.58%).

Table 2 Table show the validation results of models with different parameter setting used in training.

Model(s)	Simple/advance	Best model (check/uncheck)	Training/ testing ratio	Probability threshold	Text normalization model	Feature ranking algorithm	Overall Precision	Positive Precision	Negative Precision
Model 1	Simple	Uncheck	80/20	Default (0.50)	Default (Smoothed Relative Frequency)	Risk Ratio	91.67%	95.83%	87.50%
Model 2	Simple	Checked	80/20	Default (0.50)	Smoothed Relative Frequency	Risk Ratio	89.58%	95.83%	83.33%
Model 3	Advanced	-	70/30	Default (0.50)	Smoothed Relative Frequency	Risk Ratio	86.11%	86.11%	86.11%
Model 4	Advanced	-	90/10	Default (0.50)	Smoothed Relative Frequency	No Feature Ranking	91.67%	100.00 %	83.33%
Model 5	Advanced	-	90/10	(0.30)	Smoothed Relative Frequency	No Feature Ranking	91.67%	100.00 %	83.33%
Model 6	Advanced	-	90/10	(0.80)	Smoothed Relative Frequency	No Feature Ranking	91.67%	100.00 %	83.33%
Model 7	Advanced	-	80/20	Default (0.50)	Okapi BM25	Risk Ratio	91.67%	95.83%	87.50%

We can see that the validation phase shows exact similar results. The best model still performs well with same overall, positive, and negative precisions. So, we can consider this model for our final testing phase.

The final test results show overall precision of 69.57%, positive precision of 65.22%, and negative precision of 73.91%.

Comparing the results obtained in SVM classifier model, we got the overall accuracy as 83.33%, positive accuracy as 91.66%, and the negative accuracy as 75%. We can see the difference in both models from SVM an SAS have a huge difference. SVM model performs better in predicting the sentiment of the same dataset. So, we can say that SVM is a better model for predicting sentiments in apple reviews dataset.

4. Conclusion

We can see from the analysis in R studio that the some of the emotions like joy, anticipation are more prominent whereas surprise we expressed the least in the reviews for Apple products. Also, the top five words in each category showed which words are more relevant in the data set for a specific emotion. This can tell us how we can respond to user review in a better way. Although SVM model performed better as compared to the model created in SAS Sentiment Analysis studio, both models were able to predict the positive and negative sentiments in the reviews.

5. References

1. Package 'caret. (2021). [E-book]. In *Package 'caret* (Version 6.0-88 ed., pp. 1–224).

6. Appendices

1. Screen shots of the results window of model one testing phase.

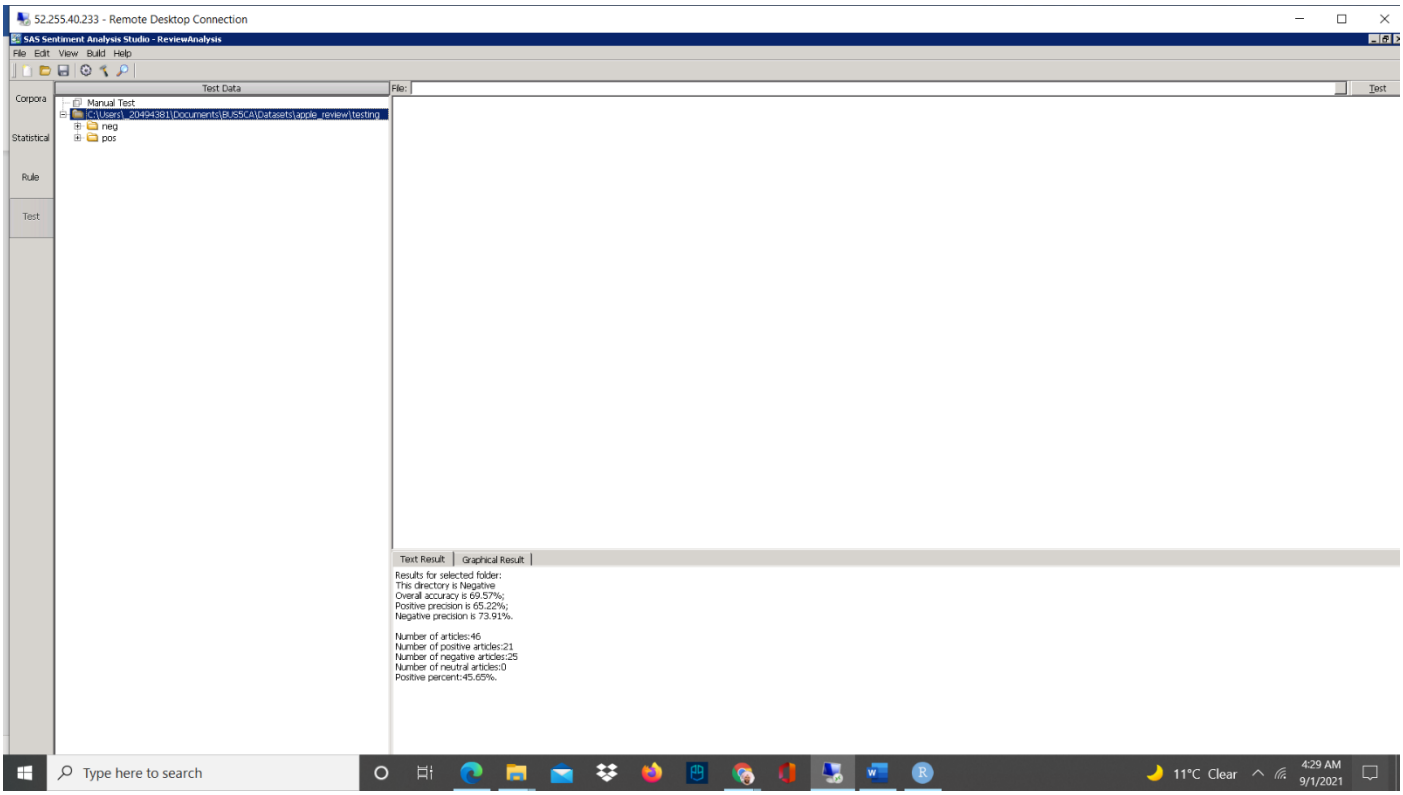

Figure 6 Model one testing results.

2. Plotting the sentiment (negative or positive).

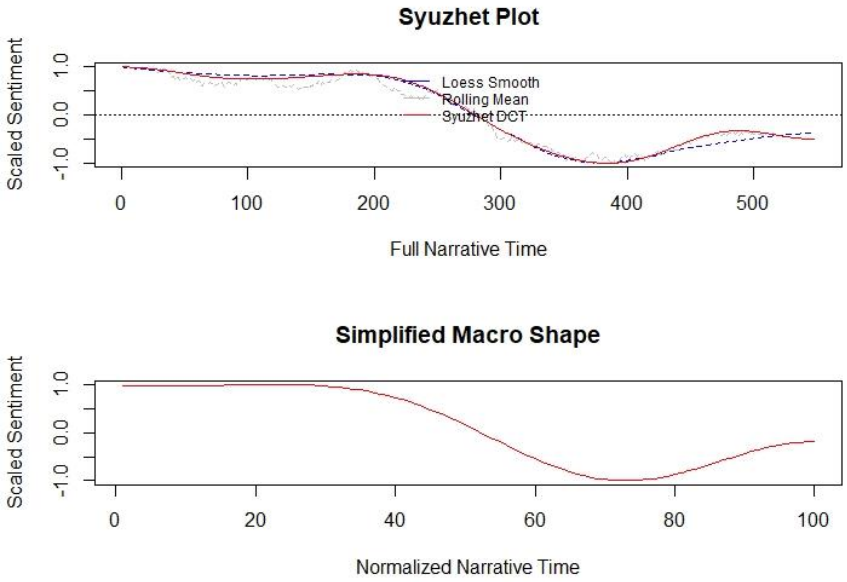

Figure7 Sentiment plot with smoothing methods applied.