Assignment 2

CS 6320.002: Natural Language Processing
Name: Manas Bundele
MMB180005

Part 1:

Writeup Question 1.1)

We tagged the words with EDIT_ because we didn't want to lose the clarification given by the editor for the text and also, at the same time, keep a way to differentiate what was written by the author and changes made by the editor so that we do not mistakenly assume something that was not written by the author(but clarified by editor) as something written by author.

If we would have deleted those words completely, we would have lost the clarification of the text that could help in better understanding of the text to our model. The text in square brackets help connect the dots in some ways.

If we would have removed the square brackets but left the words untagged, we would not have an idea of what is written by the author and what changes are made by the editor. Thus, our model might interpret something that was not meant to be understood(due to changes made by editor). Also, it would create problems if we use bigram or n-gram(n>1) classifiers.

The way we did, we can also give some less weightage to words written by editor so that our model doesn't assume something that was not said by the author but also get an idea of what the author could mean. This way the square brackets wont be able to affect the sentiment of the sentence to a great extent.

Writeup Question 1.2)

I used a set to store the negation ending tokens as it will be an O(1) search for the word instead of O(n) for a list, where n refers to number fo negation ending tokens. This will play a big role as we check it for each word once we get a negation starting word.

Part 3:

Writeup Question 3.1)

Precision measures how many of the predicted positive are actually positive.
Recall measures how many of the actual positive are labelled as positive by our model.
F-measure is the weighted harmonic mean of precision and recall. It seems to create a balance between both precision and recall as only one of them is not enough to evaluate our model.

We need all three of them because each measure focuses on one aspect of the problem. Precision is important when the cost of having false positive is high whereas recall is important when the cost of having false negative is high. F-measure tends to balance out the two but still its important for us to get an estimate of both precision and recall as average might also mislead us in some way. We focus on F-measure when both precision and recall are equally important for us.

Writeup Question 3.2)

Precision, Recall and F-measure for Gaussian NB model is (0.6261682242990654, 0.8096676737160121, 0.7061923583662714).

Writeup Question 3.3)

Precision, Recall and F-measure for LogisticRegression is (0.7647058823529411, 0.7069486404833837, 0.7346938775510203).

Logistic regression performed better on this data than Naïve Bayes. If we compare the precision of the two models, it is better for LR that means that it detected less false positives(negative sentences detected with positive sentiment) than NB. If we compare the recall, it is better for NB that means that it detected less false negatives(positive sentences detected with negative sentiment) than LR. As both the cases are equally important to us(assuming true positives are mostly similar in number for the two cases for understanding), we compare F-measure value and since its higher for LR, LR performed better on the data.

Writeup Question 3.4)

Top 10 features with weights of the LR model: (weights are sorted by absolute weight but displayed by actual value)

[('too', -3.404715203401824), ('bad', -2.378445457612999), ('dull', -2.1142058863959696), ('fails', -1.7892877344342875), ('boring', -1.7525039820661696), ('still', 1.7328961399776632), ('best', 1.5774015838982247), ('enjoyable', 1.5113324690775822), ('entertaining', 1.4395212582378303), ('and', 1.393859459363033)]

We sort them by absolute value of weight because the high positive values are important for detection of positive sentiment whereas high negative values are important for detection of negative sentiment. Positive values give more weight to positive features whereas negative values assign more weight to negative features. Thus, if we have both positive and negative features, a weighted sum decides which way the sentiment is going.

Part 4:

Writeup Question 4.1)

Performance of the new model in terms of precision, recall and f-measure: (0.7267605633802817, 0.7794561933534743, 0.7521865889212829)

Its top 10 features are(sorted by absolute values):
[('too', -3.4217570171226828), ('dal_pleasantness', 3.107017783720219), ('bad', -2.306436396810832), ('dull', -2.0219705981500136), ('fails', -1.7515827120751493), ('still', 1.7448187196369427), ('boring', -1.7383736447265503), ('best', 1.5246490053254607), ('enjoyable', 1.4995021176386523), ('entertaining', 1.4304322784701695)]

The pleasantness(evaluation) value has come up into the top 10 features for detecting positive sentiment. So, more false positives have been detected thus less precision than before. More recall shows that less false negative values and more true positives have been detected than before. Since we have seen in part 3.3, that we can focus on F-measure more as we care about both false positive and false negative here, the F-measure has improved for this model.

Part 5:

Writeup Question 5.1) 2 days excluding extra credit and writeup questions

Writeup Question 5.2) Yes, discussed a few ambiguities in the part 2 of the assignment related formation of feature dictionary. Also, about the formation of regex for different cases of single quotes.

Part 6:

Writeup Question 6.1)

The precision, recall and f-measure values are: (0.8150943396226416, 0.6525679758308157, 0.7248322147651006)

The top 10 features are (sorted by absolute values):

[('dal_pleasantness', 4.2154551020674305), ('too', -3.4034328396685827), ('bad', -2.273762855791178), ('dull', -1.98601325379383), ('fails', -1.7415064724402842), ('still', 1.7190117310015205), ('boring', -1.6164050679716024), ('best', 1.4768045206384477), ('NOT_still', 1.4048677238454654), ('entertaining', 1.4010373445108126)]

Pleasantness (evaluation) metric has become the most important feature for feature detection in this version. Few other feature weights have changed like ‘NOT_still’ has come in the top 10.

When compared with the old version of score_snippet, the precision value has increased, recall has decreased and F-measure has decreased. There might have been a change in the number of true positives but overall false positives seem to have decreased whereas false negatives have increased affecting precision and recall. Overall DAL metric values for a snippet also depends on which synonym or antonym we selected first and what was its value for all 3 metrics, activeness, pleasantness and imagery.