NLP Homework-2

Part-1

Question 1.1 :

We handled the square brackets the way we did because we wanted a proper part of speech tagging for all the words. By adding EDIT_ tags to all the words in the square brackets, we could get clarity on the meaning of the sentence and thus the part of speech tagging could be done correctly which In turn helped in building the models correctly.

The pros and cons of each method are described below :

Method 1 : Removing the square brackets and adding EDIT_ tags to the words in the square brackets

Pro : This method helps in part of speech tagging as the words added by the editor clarify the meaning of the snippet and therefore, complete a sentence which may otherwise sound incomplete. In this way, the words can be tagged with their correct parts of speech, which in turn helps in understanding the sentiment of the snippet better.

Con : This method increases the computational time as a lot of time is wasted in searching for the square brackets in each snippet and tagging the words in the square brackets. Further, the edit tags would have to removed for part of speech tagging and reattached after it is done.

Method 2 : Deleting the words completely

Pro : This method would save time as we do not need to remove and reattach the edit tags. Further we also do not need to worry about the edit tags when we are adding words to the vocabulary or when we are checking for words in the test set.

Con : This method may lead to incorrect part of speech tagging as some sentences of the author may not be complete or meaningful without the insights of the editor. This incorrect tagging may also lead to incorrect sentiment labelling of the snippet.

Method 3 : Removing the square brackets and leaving the words untagged

Pro : This method reduces the computation time as we do not need to remove and reattach edit tags for part of speech tagging.

Con : This method adds more features to the model. The features added do not represent the writing of the author as they are insights provided by the editor. These newly added features are irrelevant and can be misleading from the perspective of sentiment analysis as these features represent the editor and not the author and can thus lead to incorrect labelling of test snippets.

Question 1.2: I used the 'set' datatype to save the negation-ending tokens. I used this data type because the set of negation ending tokens is very small and each of these tokens is unique. Further, the values of sets are pre hashed so, when we want to find out if a word is present in a set or not the result is produced faster. In other words, the look up time for a set is less when compared to other data types.

Part-3

Question 3.1 : Precision measures the ratio of the number of instances that actually belong to the positive class that were identified by the algorithm to the total number of instances the algorithm labelled as belonging to the positive class. Recall measures the ratio of the number of instances that were correctly identified by the algorithms as belonging to the positive class to the total number of instances that belong to the positive class. The f-measure is the harmonic mean of precision and accuracy. All three measures are important as they give us different aspects of the data. Precision takes into account the false positives that is those instances that actually belong to the negative class but have been wrongly labelled as belonging to the positive class by the algorithm, so precision gives us the exactness of the classifier. Recall takes into account false negatives that is those instances that actually belong to the positive class but have been wrongly labelled as belonging to the negative class by the algorithm, so recall gives us the sensitivity of the algorithm. F-measure gives us a metric that is equivalent to accuracy for an unbalanced dataset and this measure gives equal importance to both precision and recall but weighs the lower of the two numbers more heavily.

Question 3.2 : The performance of the GaussianNB model is :

Precision : 0.6276346604215457

Recall : 0.8096676737160121,

F-measure : 0.7071240105540897

Question 3.3 : The performance of the LogisticRegression model is

Precision : 0.7746031746031746

Recall : 0.7371601208459214

F-measure : 0.7554179566563468

The LogisticRegression model performed better on the given data. I think this might be because logistic regression models tend to perform better when the dataset is large. Further, naïve bayes being a generative model makes the conditional independence assumption which means that the features are independent of each other which may not be an accurate assumption in the case of sentiment analysis of a document as the features, which in this case are words, are

actually dependent on each other whereas the logistic regression model is a discriminative model which does not make any such assumptions.

Question 3.4 : The top 10 features of the Logistic Regression model along with their weights are :

1. too - 3.3568227482529536

2. bad -2.4245235204010522

3. dull -2.1189405022990133

4. still 1.8120472833956167

5. boring -1.792435902936386

6. fails -1.7779048934231099

7. best 1.6023298208999541

8. enjoyable 1.4739748738189271

9. brilliant 1.4112253940140778

10. worth 1.4033794776644077

We sort by the absolute weight, rather than the actual weight because absolute value of weight gives the importance of the feature while the sign only indicates which class the feature is associated with, that is the sign is positive if the feature is associated with the positive class and is negative if the feature is associated with the negative class. Therefore, it is important to sort by the absolute weight rather than the actual value as the absolute value gives the significance of the feature (the greater the absolute value of the weight, the greater the significance of the feature).

Part-4

Note : In order to run this part of the assignment please uncomment and comment the code sections in the get_features function as well as the main function

Question 4.1 : The performance of the new model with extra features is :

Precision : 0.7015706806282722

Recall : 0.8096676737160121

F-measure: 0.7517531556802245

The top 10 features of this model along with their weights are :

1. too -3.380471158574542

2. lexicon_pleasantness 3.3518549677605045

3. bad -2.3528277151139707

4. dull -2.0234879739861147

5. still 1.8185286713862419

6. boring -1.7781658893498093

7. fails -1.7399544369081887

8. best 1.5458589956569004

9. enjoyable 1.4600735929454889

10. way 1.4030763266346244

When compared to the old model the precision has decreased, however the recall has increased. Further, the f-measure almost remains the same(75.54,75.17). This means that the number of false positives increased while the number of false negatives increased for the new model. Further, the evaluation score has become one of the top 10 significant features. This might be because the features activeness, pleasantness and imagery are rarely 0 unlike the other features and thus have a significant impact on the logistic regression algorithm. Further, these features have helped improve the sensitivity of the algorithm. However, the impact of these features is not very large, because adding 3 features to an already large feature set would not make much of a difference to the algorithm. Further, since all of the words of the snippet were not given scores, the scores were also not accurate for the snippets, which in turn reduced the impact of these features.

Part-5

Question 5.1 : This homework took me about 30 hours to complete.

Question 5.2 : No, I did not discuss this homework with anyone.