Text Classification Assignment

You may use any programming language or tool(s) to do this assignment. Please submit the following documents:

1. A PDF report answering the questions mentioned below
2. Your code and README file in a tarball archive

Deadline: 5:30pm on Sun, Oct 29

Reference: Chapter4 from Jurafsky and Martin (J&M henceforth):

https://courses.iiit.ac.in/tokenpluginfile.php/c59c33a10c086ed71473df4937d1ff34/160053/mod_forum/post/85783/chap4-jm-3ed.pdf

Download the movie reviews dataset from the link:

https://iiitaphyd-my.sharepoint.com/personal/yash_a_research_iiit_ac_in/_layouts/15/onedrive.aspx?ga=1&id=%2Fpersonal%2Fyash%5Fa%5Fresearch%5Fiiit%5Fac%5Fin%2FDocuments%2FCL2%2DTA%2FMovie%2Dreview%2Ddataset

Task: Classifying movie reviews into “postive” or “negative” classes using the algorithm mentioned in Figure 4.2 of J&M Chapter 4. Then submit a report after performing the following operations on the dataset you downloaded:

1. (30 points) Train separate naive Bayes classifiers on the training set (5 points per classifier):

i) Bag of words method using word frequencies
ii) Bag of words method using word frequency as 1 (i.e. after binarization)
iii) Content word frequencies (ignore function words)
iv) Content word frequencies of 1 per word (ignoring function words after binarization)
v) Bag of words method using word frequencies after applying the negation feature
vi) Bag of words method using word frequency as 1 (i.e. after binarization) after applying the negation feature

Negation feature: Prepend the prefix NOT to every word after a token of logical negation (*n’t, not, no, never*) until the next punctuation mark. Thus the phrase:

“didn’t like this movie , but I” becomes “didn’t NOT_like NOT_this NOT_movie , but I”

Note: You can get a list of English function words from NLTK at the end of this assignment (turn overleaf)

2. (5 points) Run each classifier above on the test set and create a confusion matrix along with a

separate table denoting precision, recall, accuracy and F1 score.

3. (5 points) Write a short note on the cases where your system misclassified sentences. Please use inguistic examples and highlight particular features to illustrate your points.

Extra credit (10 points): Incorporate features based on a standard polarity lexicon of your choice for the task of sentiment analysis and report the results. Some popular resources:

1. General Inquirer (Stone et al., 1966) 2. LIWC (Pennebaker et al., 2007) 3. Opinion lexicon of Hu and Liu (2004) 4. MPQA Subjectivity Lexicon (Wilson et al., 2005).

List of English function/stop words (Courtesy: NLTK toolkit**)**

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
"you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',
'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's",
'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what',
'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is',
'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about',
'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above',
'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under',
'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why',
'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some',
'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now',
'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn',
"couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn',
"hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't",
'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't",
'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```