Task

Classify movie reviews into positive and negative. Classification task should be done using two different classification methods (e.g. logistic regression and Naive Bayes)
Requirements
Use Python programming language
Try to achieve at least 75 - 80 % accuracy.
Dataset
http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.tar.gz

Questions

1. Describe text processing pipeline you have selected.

Before selecting the pipeline an statistical analysis of the data was carried out, to know the frequency distribution of the most used words and the less used ones. The pipeline is the same for both algorithms except for the use of a tokenzer for the NB (before its execution)

1.-Remove punctuation: With the information about the most used punctuation elements in the text, those are filtered.
2.-Remove the least used words: Those words that are used less than 6 times in the whole text are marked and removed.
3.-Remove the Stopwords: A personalized file is used in order to remove the Stopwords.
4.-Lemmatization: Obtain the lemmas for every of the final words.
5.-Tokenize words for each classifier: Each classifier uses its own tokenizer

2. Why you have selected these two classification methods?

Multinominal Naive bayes (NB) was chosen due to its popularity in the field after reading some papers and documentation, I realized it was fast and quite accurate. It also outperformed several of the other algorithms I tried and it was easy to tune.
On the other hand the chosen CNN, might seem more time consuming, but thanks to the used of an embedding layer and well adjusted hyperparameters through random search, it ends outperforming most of the other options. Finally, at first glance, the main reason to go with it was, undoubtedly, its flexibility and adaptability to any problem (if well adjusted).

3. Compare selected classification methods. Which one is better? Why?

Both classifiers perform almost the same (86,41% CNN, 86,65% NB), I guess, even if CNN is really flexible it is a tough classifier to tune and perhaps with a different layer distribution or more hyperparameters (in the choosing dictionary), it could clearly outperform the NB. On the other hand, NB is, as seen in lots of papers, the most popular option in document analysis, as it can obtain good results in a decent time while being easy to adjust.
To sum it up, the NB is better, due to its fine tunning of hyperparameters and normally good performance in this field.

4. How would you compare selected classification methods if the dataset was imbalanced?
Results

In the case of an imbalanced dataset, a popular solution (the one I would choose) is to balance it artificially, do an statistical analysis of the frequency of all the words in both datasets and eliminate the least used ones by setting a threshold that would remove more words in a dataset, so that balance is achieved.

Everything (source code, answers to questions, etc.) should be packed into single github repository.