Report – Natural Language Processing (CSCE 689 – sentiment Analyzer)

Name**: Shiva Kumar Pentyala**
UIN: **127003995**

Go through the README file for compilation related instructions.

Q1. Multinomial Naive Bayes classifier:

Output -

[INFO] Fold 0 Accuracy: 0.765000
[INFO] Fold 1 Accuracy: 0.850000
[INFO] Fold 2 Accuracy: 0.835000
[INFO] Fold 3 Accuracy: 0.825000
[INFO] Fold 4 Accuracy: 0.815000
[INFO] Fold 5 Accuracy: 0.820000
[INFO] Fold 6 Accuracy: 0.835000
[INFO] Fold 7 Accuracy: 0.825000
[INFO] Fold 8 Accuracy: 0.755000
[INFO] Fold 9 Accuracy: 0.840000
[INFO] Accuracy: 0.816500

Analysis - This version of Naive Bayes gave the highest accuracy.

Q2. After removing the Stop words:

Output -

[INFO] Fold 0 Accuracy: 0.765000
[INFO] Fold 1 Accuracy: 0.825000
[INFO] Fold 2 Accuracy: 0.815000
[INFO] Fold 3 Accuracy: 0.830000
[INFO] Fold 4 Accuracy: 0.795000
[INFO] Fold 5 Accuracy: 0.830000
[INFO] Fold 6 Accuracy: 0.835000
[INFO] Fold 7 Accuracy: 0.835000
[INFO] Fold 8 Accuracy: 0.760000
[INFO] Fold 9 Accuracy: 0.820000
[INFO] Accuracy: 0.811000

Analysis – Accuracy in this case is lower than that of the previous version. In some applications removing all stop words right from determiners (e.g. the, a, an) to prepositions (e.g. above, across, before) to some adjectives (e.g. good, nice) can be an appropriate stop word list. To some applications however, this can be detrimental. For instance, in our case (sentiment analysis) removing adjective terms such as ‘good’ and ‘nice’ as well as negations such as ‘not’ can throw algorithm completely off its track. In order to compensate, we can choose to use a minimal stop list consisting of just determiners or determiners with prepositions or just coordinating conjunctions depending on our need of application.

Q3. Binary version of Naive Bayes Model:

Output-

[INFO] Fold 0 Accuracy: 0.690000
[INFO] Fold 1 Accuracy: 0.670000
[INFO] Fold 2 Accuracy: 0.755000
[INFO] Fold 3 Accuracy: 0.735000
[INFO] Fold 4 Accuracy: 0.695000
[INFO] Fold 5 Accuracy: 0.735000
[INFO] Fold 6 Accuracy: 0.755000
[INFO] Fold 7 Accuracy: 0.700000
[INFO] Fold 8 Accuracy: 0.730000
[INFO] Fold 9 Accuracy: 0.715000
[INFO] Accuracy: 0.718000

Analysis - This performed badly in comparison to the above versions because of the simplification in the model assumptions.

Q4. I think considering below features may improve our model accuracy and many of these features are independent of the Bag of words features that we considered.

1. Features based on subjective sentence occurrence statistics
2. delta-tf-idf weighting of word polarities
3. sentence-level features based on polarity. We can exploit the structure in sentences, rather than seeing a review as a bag of words. For instance, conjunctions were analyzed to obtain the polarities of the words that are connected with the conjunct.
4. Occurrence of subjective words
5. Punctuations like Exclamation and Question marks
6. The first line, last line polarity can also be used as a feature because generally first and last lines of a review are often very indicative of the review polarity.
7. Emotion icons
8. Negation words
9. Intensity words(very,really etc)
10. elongated words (eg. goooood)
11. unigrams and bigrams of every word (particularly if you have a large corpus)

Problems and Limitations:

1. Naive Bayes is based on Bag of Words Model. So, it doesn't classify correctly in some cases.
2. In Binomial version of Naive Bayes model, we calculate probability in terms of number of documents, which may be super simplified assumption in some of the documents.
3. Naive Bayes classifier may make a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class. Due to this, the result can be potentially bad.

4. Another problem may happen due to Data Scarcity. For any possible value of a feature, we need to estimate a likelihood value by a frequentist approach. This can result in probabilities going towards 0 or 1, which in turn leads to numerical instabilities and worse results. In this case, we need to smooth in some way your probabilities or to impose some prior on our data.
5. A third problem may arise for continuous features. It is common to use a binning procedure to make them discrete, but if we are not careful we can throw away a lot of information. Another possibility is to use Gaussian distributions for the likelihoods.

Sources: Wikipedia, IEEE, quora.

Thank you