Sentiment Analysis Using Naïve Bayes Text Classification

Objective: The main goal of this assignment to classify IMDb movies reviews either positive or negative by using naïve bayes which is built from scratch.

Preprocessing: The preprocessing contains four different scenarios.

1. No preprocessing: In this just the basic symbol cleaning is done.
2. With Lemmatization: In this the words are reduced to their root words.
3. With lemmatization and stop words removal: In this the words are reduced to their root words and also stop words are removed. The words like is, an, the and etc., which doesn't play crucial role in sentiment analysis.
4. With lemmatization, stop words removal and Handling logical negation: In this lemmatization is done along with stop words removal and also taking care of the negated sentiment.

The libraries which are required are imported and handling the missing values.

Data preparation: The data was loaded, shuffled and split into training set and testing set. The training set contains 1000 reviews and testing set contains 200 reviews (100 positive and 100 negative).

Model Building: The naïve bayes model is built from scratch

1. Calculating prior probabilities for each class.
2. Calculating likelihood of each word to given class.
3. Calculating the posterior probability using log.
4. Laplace smoothing is done to ensure zero probability.
5. Removing unknown words.

Performing sentiment analysis on the test data.

Model Evaluation: In this the model is evaluated.

After conducting sentiment analysis on the test data the model performance is evaluated by different metric precision, recall and F1 score. And also the creating a confusion matrix. A confusion matrix provides detailed breakdown correct and incorrect predictions.

Results: *Scenario 1 & Scenario 2* both nearly shows the same results, only lemmatization didn't add much value. The scores of the both scenarios are same.

The model performed well in *scenario 3 and scenario 4* where the stop words are removed and negation is handled. The model focused on the words which are crucial for the sentiment and bag of words problem is also solved by the logical negation.

Answers for the Questions:

1.

Scenario	Class	Precision	Recall	F1
No Preprocessing	Negative (0)	0.7008	0.8900	0.7841
	Positive (1)	0.8493	0.6200	0.7168
Lemmatization Only	Negative (0)	0.7008	0.8900	0.7841
	Positive (1)	0.8493	0.6200	0.7168
Lemmatization + Stopwords	Negative (0)	0.7120	0.8900	0.7911
	Positive (1)	0.8533	0.6400	0.7314
All Preprocessing	Negative (0)	0.7143	0.9000	0.7965
	Positive (1)	0.8649	0.6400	0.7356

The negative class is showing the best results recall (0.90) and F1(0.7965) whereas the positive class showed consistent gains in precision at 0.8649. The recall is low throughout all scenarios.

Scenario 1 & Scenario 2: Both nearly shows the same results, only lemmatization didn't add much value. Without preprocessing the model didn't perform well.

Scenario 3: The stop words are removed which made to the model to deliver better results the words like 'a', 'the',' is' and etc., doesn't play much role in sentiment analysis. As these kind of words are removed the model focused on the words which carry sentiment which results in better prediction.

Scenario 4: We have introduced logical negation along with lemmatization and removing stop words made the model to deliver better predictions. The words like not and no can completely change the sentiment of the sentence.

I suggest scenario with lemmatization, stop words removal, and handling logical negation because this combinations always provides the best results. It is the combination of all the preprocessing methods. Removal of stop words and handling logical negation helps the model in better prediction.

2. Laplace smoothing is a technique used in naïve bayes to prevent assigning zero probability to the class which is never seen in the training phase, which helps in better results while calculating conditional probability. It is done by adding a small constant alpha in the numerator and alpha times number of possible outcomes to the

denominator. It assigns a small values to the classes not seen in the training phase to make the probability non-zero, which improves the model accuracy.

The regular conditional probability calculation is

P(A|B) = count(A, B)/ count(B)

If A did not appear in class B the probability is zero.

After applying laplace smoothing

P(A|B) = count(A, B) + alpha / count(B) + alpha. Mod(n)

If we didn't use laplace smoothing, the probability of the class that is not seen will be zero even though other words strongly suggest that class. Not only that the predictions will be biased towards the class which have the words repeated most number of times and because of that the accuracy will be dropped due to the predictions are skewed by missing words.It also helps in calculating logarithmic probability because the log of zero is infinity the model will be crashed.

3. The one of the limitation of naïve bayes is the assumption of conditional independence between words which is the model assumes that presence of one word does not affect the presence of any other word in the same sentence which is not true.

a) The model cannot understand Negation and Irony

For example if we take a sentence "the movie was not good but the actions parts are wonderful." In this the model sees the positive words like good and wonderful and predict it as a positive review. It couldn't consider the not in the sentence.

b) The model sometimes misclassify because of bag of words

For example if we take a sentence "This movie is amazing, but is sequel was horrible." The model consider the words amazing and horrible which carry the most weight in sentiment and it might misclassify the review based on the probability of both the words.

c) Mixed reviews

Some of the sentences contains mixed revies the bag of words cannot understand the order of the words it predicts the output based on the probability. In case of a mixed review it might be misclassified.

d) Sarcasm

The review with a sarcasm may be misclassified and positive because the model cannot understand sarcasm right.