Compare between multinomial and multivariate models for naïve bayes classifier

The multinomial model is designed to determine term frequency i.e. the number of times a term occurs in a document. Considering the fact that a term may be pivotal in deciding the sentiment of the document, this property of this model makes it a decent choice for document classification.

In the multivariate Bernoulli Naïve Bayes Classifier algorithm, features are independent binary variables which represents that whether a term is present in the document under consideration or not. Being slightly similar to the multinomial model in the classification process, this algorithm is also a popular approach for text classification tasks but it differs from the multinomial approach in the aspect that multinomial approach takes into account the term frequencies whereas Bernoulli approach is only interested in devising that whether a term is present or absent in the document under consideration.

For the 5-fold cross-validation on training data results in terms of accuracy (in percentage) are as follows:

Folds	Multinomial	Multivariate
1st fold	84.08071748878923	84.19282511210763
2nd fold	92.152466367713	95.85201793721974
3rd fold	97.86995515695067	97.6457399103139
4th fold	98.31649831649831	97.8675645342312
5th fold	98.65319865319864	97.3063973063973
Average of all folds	94.214	94.57
Final on test data	**96.76840215439856**	**92.81867145421903**

It is concluded from this experiment that Multinomial Naïve Bayes performs slightly better than Multivariate Naïve Bayes