Naive Bayes models comparison for text classifying

Gabriele Bruni

September 2019

1 Introduction

In this work, I explore two different probabilistic models used in text classification, both of which belong to the ***naive Bayes*** method family. All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. In the ***multivariate Bernoulli*** model, features are independent binary variables describing inputs. In ***Multinomial*** model, a feature vector is a histogram, with counting the number of times each event was observed in a particular instance. The aim of this work is to better understand the difference between the two models, given different data set, using different feature selection and different vectorization methods. We find that generally, the Multinomial model's performance scales well with increasing size of the data set while the Bernoulli one does a good job classifying short text.

2 The Naive Bayes Classifier

A widely used framework for classification is provided by a simple theorem of probability known as *Bayes' rule:*

$$P(C = c_k \mid \mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x} \mid C = c_k)\, P(C = c_k)}{P(X = x)}$$

C is a random variable whose possible values are the class to which each document can belong. $\boldsymbol{X}$ is a vector random variable whose values are vector of features values $\boldsymbol{x} = (x_1, ..., x_n) = (x_j)_{j=1...n}$, one vector for each document.
$P(c_k|x)$ is the conditional probability that a document belong to class c_k.
$P(\mathbf{x}|c_k)$ can be calculated assuming that each feature x_j is *conditional indipendent given the class* c_k. Thus, given a features vector $\mathbf{x} = (x_j)$, the class conditional probability cam be calculated as

$$P(\mathbf{x}|c_k) = \prod_j P(x_j = c_k)$$

The "individual" likelihoods for every feature in the feature vector can be estimated via the maximum-likelihood estimate, which is simply a frequency in the case of categorical data:

$$P(x_j|c_k) = \frac{N_{x_j,c_k}}{N_{c_k}}$$

where:

- N_{x_j,c_k}: number of times feature x_j appears in samples from class c_k.
- N_{c_k}:total count of all features in class c_k .

2.1 Multi-variate Bernoulli Naive Bayes

The Multi-variate Bernoulli model is based on binary data: every word in the feature vector of a document is associated with the value 1 or 0. The feature vector has m dimensions where m is the number of words in the whole vocabulary ; the value 1 means that the word occurs in the particular document, and 0 means that the word does not occur in this document. The class conditional distribution can be written as

$$P(\mathbf{x}|c_k) = \prod_j P(x_j|c_k)^b (1 - P(x_j|c_k))^{(1-b)} (b = 0, 1)$$

2.2 Multinomial Naive Bayes

A alternative approach to characterize text documents — rather than binary values — is the term frequency . Define $N_{i,t}$ to be the count of the number of times word w_t occurs in document $\mathbf{x}$. The class conditional distribution can be written as

$$P(\mathbf{x}|c_k) = P(|\mathbf{x}|)|\mathbf{x}|! \prod_t \frac{P(w_t|c_k)^{N_{i,t}}}{N_{i,t}!}$$

2.3 Classification

Classification can be performed on test documents by calculating the posterior probability of each class given the evidence of the test document, and selecting the class with the highest probability.

3 The Method

We tried to classify some reviews of products from amazon, based on:

- their **score** (from **1 to 5**, five class label).
- the **sentiment** - positivity or negativity of the review (**0 or 1**, two class label).

	Toys and Games	Digital Music	Grocery and Gourmet food
Reviews	167597	61696	143337
Average length	29.0	76.0	32.0
Variance length	3284.22	10874.22	2407.76

Table 1: Comparison between different data sets.

3.1 The Data Set

The data sets have been taken from *http://jmcauley.ucsd.edu/data/amazon/*. Each data set include:

- the text of the reviews.
- a brief summary of each review - we consider this as short review.
- score of the review (from 1 to 5).

The sentiment of the review was not present: this was added manually considering as positive the reviews with score higher or equal to 3 and negative reviews with score lower than 2.
In particular, we have played with the *Digital Music* data set, the *Grocery and Gourmet Food* data set and the *Toys and Games* data set.
In table 1 you can visualize the difference between the three data sets.

3.2 Feature Selection

To each data set, has been applied the same work of feature selection to reduce the vocabulary size. In particular, we have:

- removed all non-letter characters.
- converted all the word of the reviews to lowercase.
- filtered out stop-words. We use a list of stop words provided by the `NLTK library`.
- stemming all words using *Snowball* method.

The above text pre-processing work has been applied to both reviews and summary of the reviews and added manually to each data set.
At end each data set included four text columns: the original review, the original summary of the review, the cleaned text review and the cleaned summary.

3.3 Feature Extraction

We have used the *Bag of word* approach to codify text as features vector for our naive Bayes model. Normally the bag of word model was created using the uni-gram approach (one token/features = one word), but for some test (see chapter 3.4) we varied the number of n-gram and/or we excluded the less frequent word of the vocabulary (the ones that appear less than n times among all document).

3.4 The Tests Performed

We tried to evaluate how the two models behaved in relation to:

1. number of examples in the training set.
2. number of words in the vocabulary.
3. length of reviews (we have included the summary of each review for this purpose).
4. type of classification: **binary** in the case of the **sentiment** classification, **multi-value** in the case of **score** classification.

For the above mentioned tests the learning curves of the two models were drawn and compared.
In addition, some tests were performed to assess the impact that the work of feature selection and feature extraction has had on the classification process.

4 Results And Analysis

4.1 Toys And Games Data Set

Figure from 1 to 8 represent the above test performed on *Toys and Games data set*. We can draw the following conclusions:

- The Multinomial model seems to behaves generally better than the Bernoulli model especially in the case of multi values classification. **Figure 2** show that the difference between the two increases with the size of the training set in the.
- From **Figure 3**, we can see that when the number of features (words in the dictionary) is drastically reduced (we have limited it to 50 in this test) , the Bernoulli has slightly higher performance than the Multinomial, but the gap between the two is reduced as the size of the training set increases.
- **From the previous two point we can deduce that the Multinomial model benefits from larger dimensions of the training set, while the Bernoulli model does not seem to show improvements from this point of view**.

minDf	nGramRange	maxFeatuers	reviewLenght	Bernoulli score	Multinomial score
3	(1,3)	none	short	0.6253254804711718	**0.6831990080595164**
5	(1,3)	none	short	**0.6718226906385617**	0.6736205827650341

Table 2: Impact of feature extraction

- The above deduction is evident also in **Figure 4** in which we perform the multi value classification with reduced number of features (where the Bernoulli model normally behaves better): in this case the Multinomial is able to adapt his self better than the Bernoulli model and overcome the latter slightly. Moreover here the Bernoulli seems to have a decline in performance when increasing training set size.
- When the length of reviews is short (**Figure 5-8**), both models generally offer the same performance. However, even when few words are available in the dictionary, the Bernoulli seems to have a slight advantage over the Multinomial, but that's is not statistically relevant.
- **Hence we can deduce that Bernoulli behaves slightly better than Multinomial when the number of the features is very low (only 50 features was used to perform the test**.

4.2 Digital Music Data Set

Figure from 9 to 16 represent the above test performed on *Digital Music data set*. We can draw the following conclusions:

- We note in the case of multi value classification that the complessive accuracy score is lower. This is probably due to the major average length of the reviews compared to the previous data set.
- From **Figure 11-12** we see that **when decreasing the number of features, the Bernoulli model in this case does not perform better than the Multinomial model**. Indeed, it behaves much worse and the difference between the two widens as the size of the data set increases. Once again in support of the fact that the Multinomial model is the only one of the two that receives advantages from the larger dimensions of the training set **and that Multinomial model perform better with very long review** . In fact the length of the reviews in this data set is about doubled compared to precedent data set.
- From **Figure 13-16** we see that when using short reviews, the Bernoulli model seems to behave slightly better than the Multinomial.**To confirm the previous point, the performance of the Bernoulli model declines with the length of the reviews**
- Also in this data set, when we perform multi values classification, the Multinomial responds better.
- When the reviews are short and the number of words in the dictionary is low, both models behave the same way, and we don't have anymore the slight advantage of the Bernoulli model, to confirm the point that the Multinomial perform better on long reviews.

4.3 Grocery And Gourmet Food Data Set

Figure from 17 to 24 represent the above test performed on *Grocery and Gourmet food datataset*. We can draw the following conclusions:

- Also in this case, the Multinomial model can benefit from the size of the training test, and this allows it to work better even when the number of targets grows.
- When the number of features is limited, the Bernoulli model returns to behave like the first data set. This can be explained by the fact that the two data sets are very similar in term of length of the reviews, and testifies how the length of the reviews plays a decisive role in the increase/decrease of the performance of the Bernoulli model.

4.4 Impact Of Feature Selection

By performing all the above tests but with the non-processed data sets (no feature selection) we can say that the feature selection work does not always lead to an increase in performance: normally there is a slight improvement, especially when the number of words in the dictionary is low (**Table 4**), but when the reviews are short it turns out that the preprocessing job leads to a negative result in terms of performance (**Table 5-6**).
We report only the results for the **score classification** in **Table 3-6**.
We obtain similar results with the binary classification, but the differences are less marked and therefore less visible.

	Preprocessed	Original
Bernoulli	0.58848	0.57920
Multinomial	0.65244	0.65107

Table 3: Impact of feature selection.

	Preprocessed	Original
Bernoulli	0.58424	0.54725
Multinomial	0.60796	0.60605

Table 4: Impact of feature selection reducing number of features to 50.

	Preprocessed	Original
Bernoulli	0.64534	0.65065
Multinomial	0.64818	0.65874

Table 5: Impact of feature selection using shorter reviews (summaries)

	Preprocessed	Original
Bernoulli	0.61584	0.63645
Multinomial	0.61318	0.63430

Table 6: Impact of feature selection using shorter reviews (summaries) and reducing number of word to 50.

4.5 Impact Of Features Extraction

Table 2 shows the best result obtaneid in **score classification** with both models playing with the parameter of the feature extraction. These parameters are:

- minDf: minimum number of documents in which a word must appear to be included in the dictionary
- nGramRange: set the range of the ngram with which we decide to extract the features.
- maxFeatures: maximum number of words(features) to be included in dictionary.
- reviewLenght: type of the review use to perform the score classification (summaries = short, reviews = long).

5 Conclusions and Future Work

From the results above it is clear how the choice of the model to be used depends entirely on the type of data you want to classify and the type of classification you want to operate.
Both models work well on binary classification tasks, with the Bernoulli model slightly ahead.
In the case of classifications with multiple values, the Multinomial model is always preferable.

Exception to this general rule is the case in which the number of words in the dictionary is low **and the reviews are sufficiently short: this is the only case in which the Bernoulli model seems to offer slightly higher performances on multi-value classification task**.
When the number of word in vocabulary is restricted we note a small improvement in Bernoulli model and a small deterioration in Multinomial model, in agreement with what has been done in [2]: empirical comparisons provide evidence that the Multinomial model tends to outperform the multi-variate Bernoulli model if the vocabulary size is relatively large.
The Multinomial model perform better with long review.

Furthermore, when the training set consists of a few samples, the Bernoulli model achieves better results, but fails to benefit from the increasing size of the training set: **on large training set is usually preferable the Multinomial model.**

We have also show that in text classification, large differences in performance can be attributed to the choices of stop word removal, stemming, token-length, n-grams range, number of features, choice of the words to be included in the vocabulary, See [1] for most complete tests.
In this work the **variance** of the review lengths has been omitted, but in future work it might be interesting to evaluate its impact on the performance of the two models. Also, can be interesting evaluating the difference between lemming and stemming in text pre-processing.

References

[1] Lawrence M Rudner and Tahung Liang. *Automated essay scoring using bayes' theorem.* The Journal of Technology, Learning and Assessment, 2002.

[2] Andrew McCallum, Kamal Nigam, et al. . *A comparison of event models for naive bayes text classification.* The Journal of Technology, Learning and Assessment, 2002.

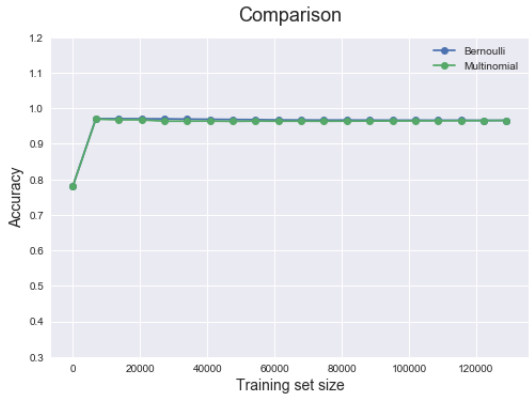

Figure 1: Comparison of both models on *Toys and Games datasets* for sentiment classification using cleaned review.

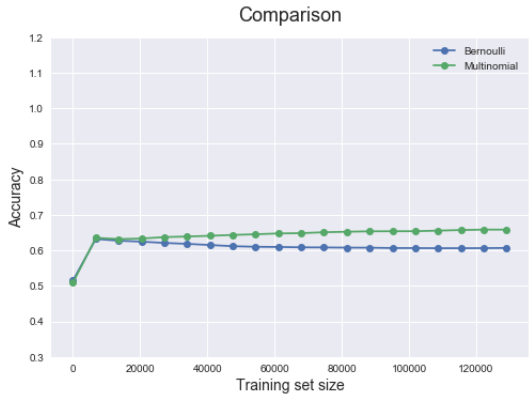

Figure 2: Comparison of both models on *Toys and Games datasets* for score classification using cleaned review.

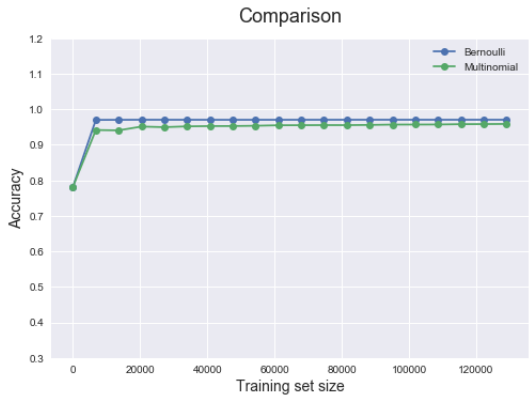

Figure 3: Comparison of both models on *Toys and Games datasets* for sentiment classification using cleaned review and only 50 words in dictionary.

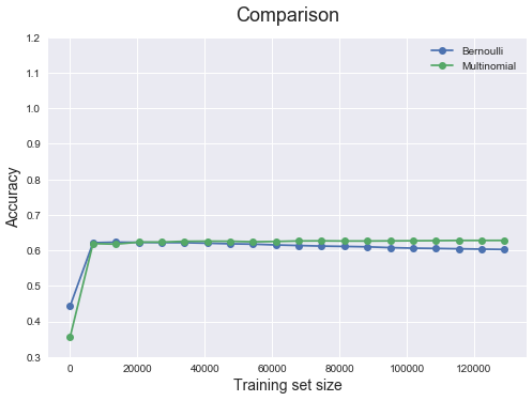

Figure 4: Comparison of both models on *Toys and Games datasets* for score classification using cleaned review and only 50 words in dictionary.

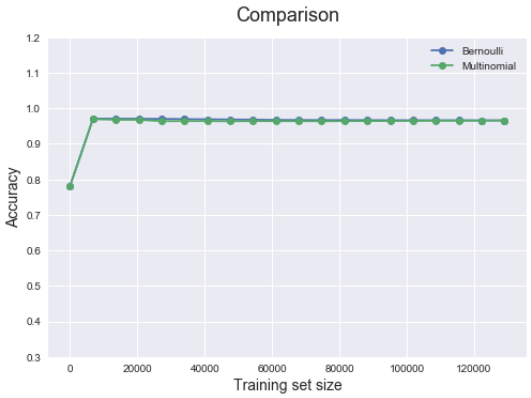

Figure 5: Comparison of both models on *Toys and Games datasets* for sentiment classification using cleaned summary reviews.

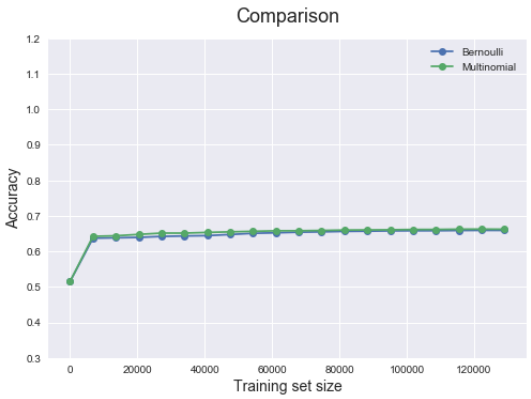

Figure 6: Comparison of both models on *Toys and Games datasets* for score classification using cleaned summary reviews.

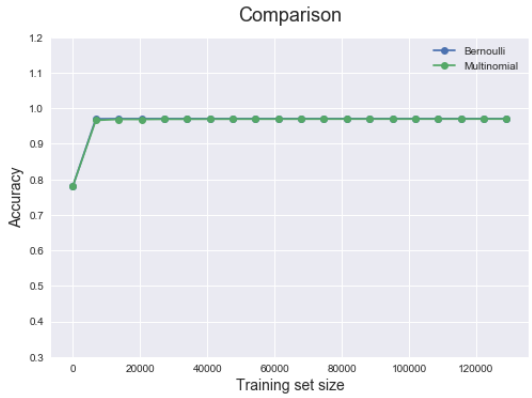

Figure 7: Comparison of both models on *Toys and Games datasets* for sentiment classification using cleaned summary reviews and only 50 words in dictionary.

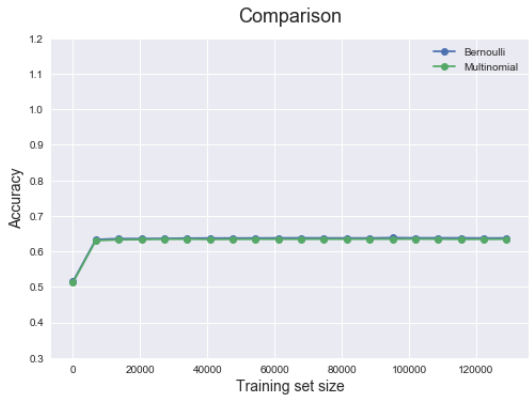

Figure 8: Comparison of both models on *Toys and Games datasets* for score classification using cleaned summary reviews and only 50 words in dictionary.

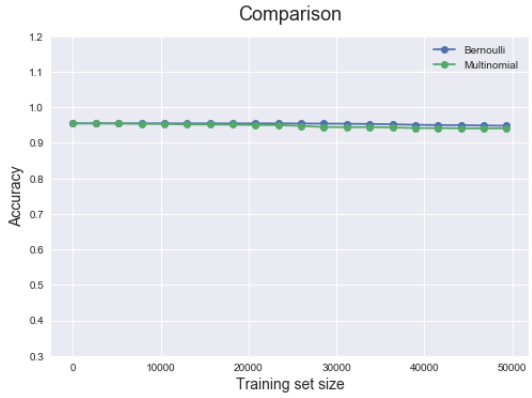

Figure 9: Comparison of both models on *Digital Music datasets* for sentiment classification using cleaned review.

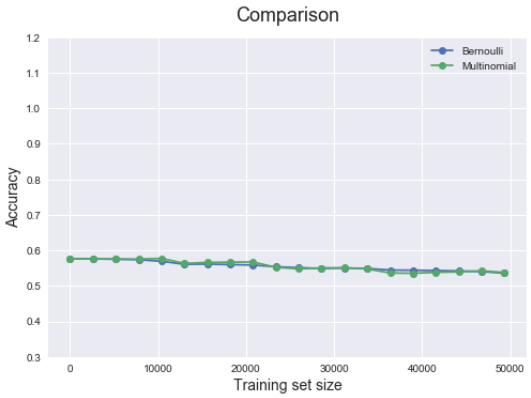

Figure 10: Comparison of both models on *Digital Music datasets* for score classification using cleaned review.

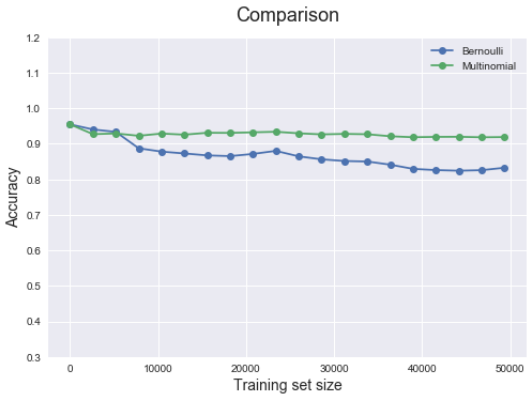

Figure 11: Comparison of both models on *Digital Music datasets* for sentiment classification using cleaned review and only 50 words in dictionary.

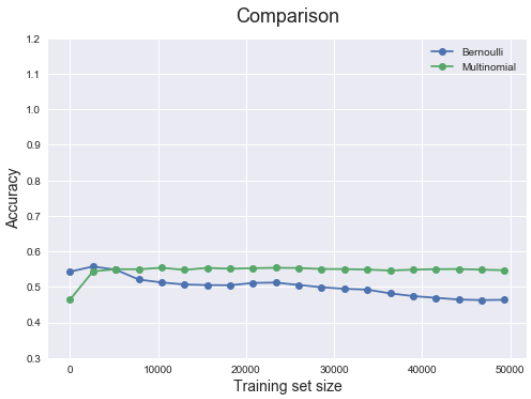

Figure 12: Comparison of both models on *Digital Music datasets* for score classification using cleaned review and only 50 words in dictionary.

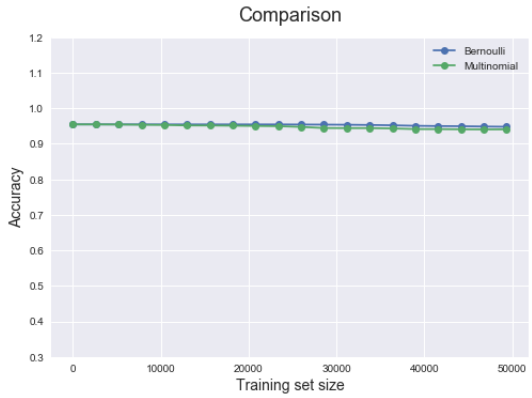

Figure 13: Comparison of both models on *Digital Music datasets* for sentiment classification using cleaned summary reviews.

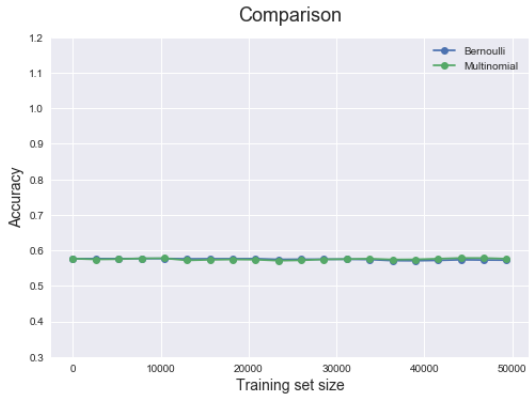

Figure 14: Comparison of both models on *Digital Music datasets* for score classification using cleaned summary reviews.

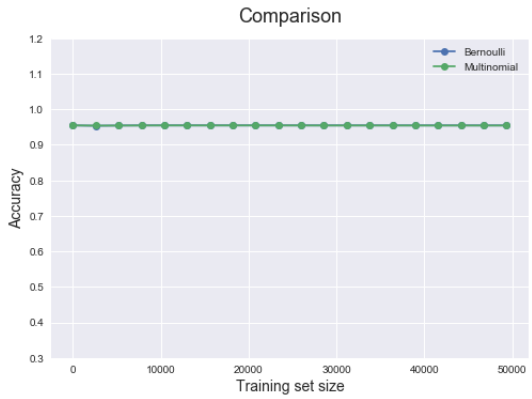

Figure 15: Comparison of both models on *Digital Music datasets* for sentiment classification using cleaned summary reviews and only 50 words in dictionary.

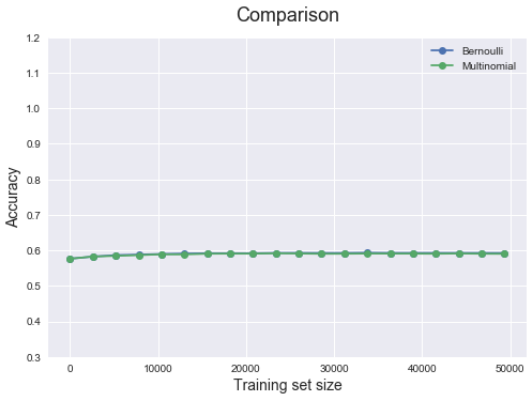

Figure 16: Comparison of both models on *Digital Music datasets* for score classification using cleaned summary reviews and only 50 words in dictionary.

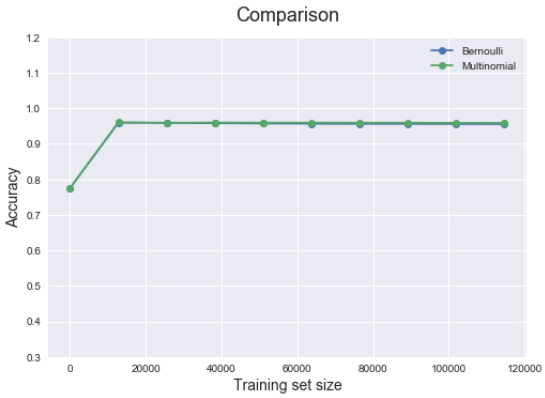

Figure 17: Comparison of both models on *Grocery and Gourmet food data set* for sentiment classification using cleaned review.

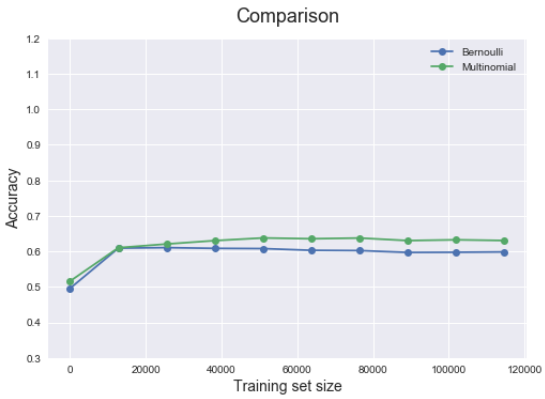

Figure 18: Comparison of both models on *Grocery and Gourmet food data set* for score classification using cleaned review.

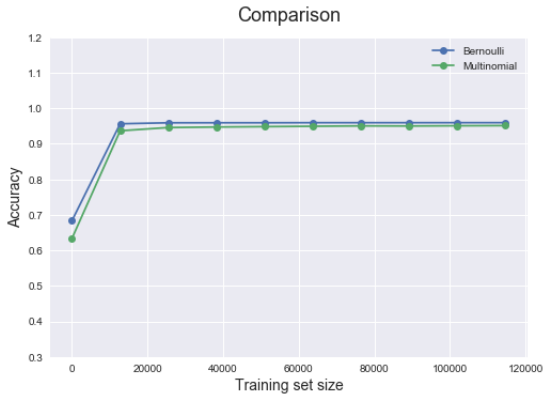

Figure 19: Comparison of both models on *Grocery and Gourmet food data set* for sentiment classification using cleaned review and only 50 words in dictionary.

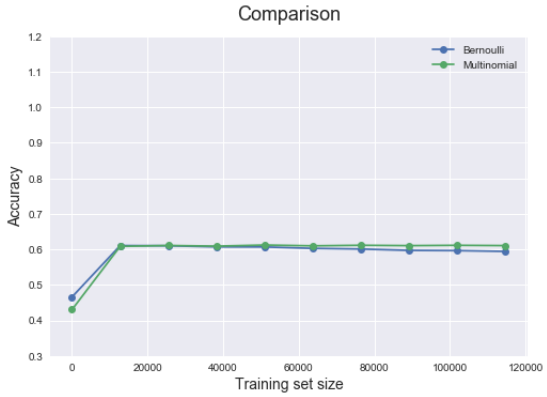

Figure 20: Comparison of both models on *Grocery and Gourmet food data set* for score classification using cleaned review and only 50 words in dictionary.

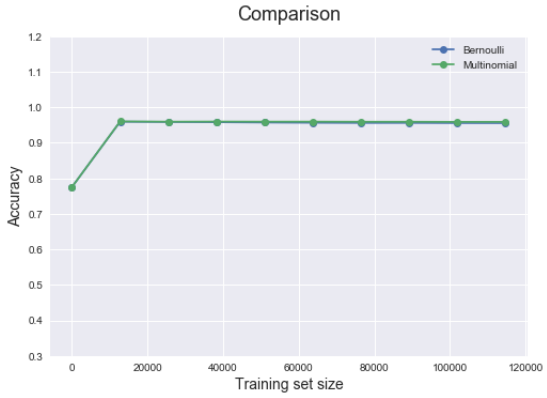

Figure 21: Comparison of both models on *Grocery and Gourmet food data set* for sentiment classification using cleaned summary reviews.

2

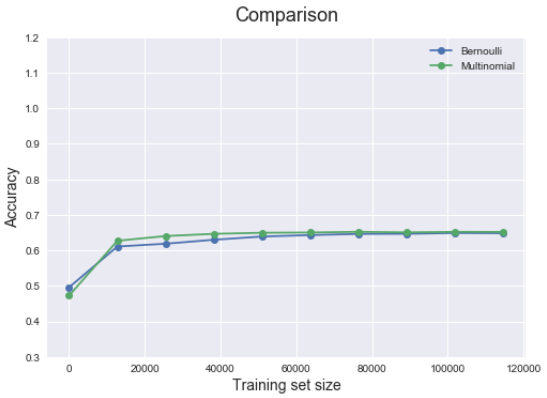

Figure 22: Comparison of both models on *Grocery and Gourmet food data aset* for score classification using cleaned summary reviews.

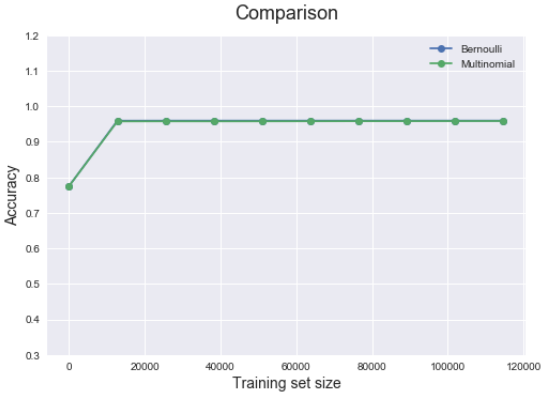

Figure 23: Comparison of both models on *Grocery and Gourmet food data set* for sentiment classification using cleaned summary reviews and only 50 words in dictionary.

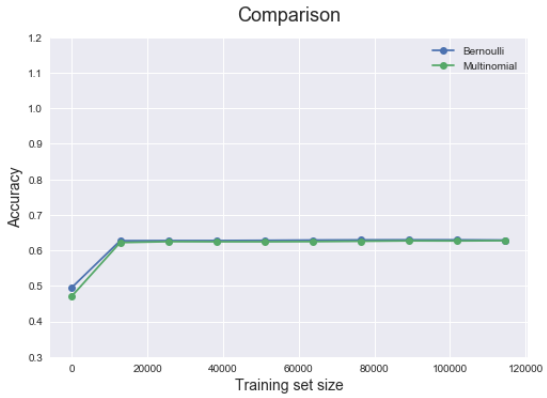

Figure 24: Comparison of both models on *Grocery and Gourmet food data set* for score classification using cleaned summary reviews and only 50 words in dictionary.