Artificial Intelligence Assignment Reports

Bruni Gabriele
6 September 2019
gabriele.bruni@stud.unifi.it

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 dataset, using different feature selection and different vectorization methods. We find that generally, the Multinomial model's performance scales well with increasing size of the dataset 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. ***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 **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** first (from **1 to 5**, five class label), then based on the **sentiment** - positive or negative of the review (**0 or 1**, two class label).

3.1 The Datasets

The datasets have been taken from *http://jmcauley.ucsd.edu/data/amazon/*, and includes reviews with relative ratings, summary of review and the helpfulness votes. 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* dataset, the *Grocery and Gourmet Food* dataset and the *Toys and Games* dataset.

3.2 Feature Selection

To each dataset, 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 *Sonowball* method.

3.3 Feature extraction

We have used the *Bag of word* approach to codify the cleaned reviews (and also the non-cleaned, see chapter 3.4) as features vector for our naive Bayes model. Normally the bag of word model was created using the unigram 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. number of class values (we must classify among 5 labels when we want to classify the score of the review, between two labels when we want to classify the feeling of the review).

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 dataset

- The multinomial model always behaves generally better than the Bernoulli model.
- **Figure 2** show that the difference between the two increases with the size of the training set.
- From **figure 3**, we can see that when the number of features (words in the dictionary) is reduced, 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 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.
- When the number of class label is high, the multinomial is always better than the bernoulli. This is made evident by **figure 4** in which at reduced number of features (where

the bernoulli model normally behaves better) is accompanied a greater number of class label (the overall field is used): 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 with high 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.

4.2 Digital Music datasets

- We note that if the number of targets is greater than 2 (when the overall field is used) the complessive accuracy score is lower. This is probably due to the major average length of the reviews compared to the previous dataset.

- From **picture 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 dataset 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.

- When using short reviews, the bernoulli model seems to behave slightly better than the binomial

- As the number of class labels increases, the performance of both models decreases, but the multinomial responds better, also taking advantage of the larger size of the training set.

- When the reviews are short and the number of words in the dictionary is low, both models behave the same way.

4.3 Grocery and Gourmet food datataset

- 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 dataset. This can be explained by the fact that the two datasets are very similar, and testifies how the length of the reviews plays a decisive role in the increase / decrease of the performance of the bernoulli model, which in case the reviews are very short, behaves exactly like the bernoulli model.

4.4 Impact of feature selection and extraction

By performing all the above tests but with the unprocessed datasets (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, but when the reviews are short it turns out that the preprocessing job leads to a negative result in terms of performance.
In the case of naive Bayes classifiers and text classification, large differences in performance can be attributed to the choice of stop word removal, stemming, and token-length (see (4) for more detail). In our tests the best result with regard to the multivalue classification was obtained by increasing the number of n-grams up to a maximum n = 3.4 and excluding those words that are not found in at least 2.3 documents.

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 have a sufficiently low average length: this is the only case in which the Bernoulli model seems to offer slightly higher performances on multi-value classification task.

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 dataset is usually

preferable the multinomial model

In this work the viability 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.

References

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

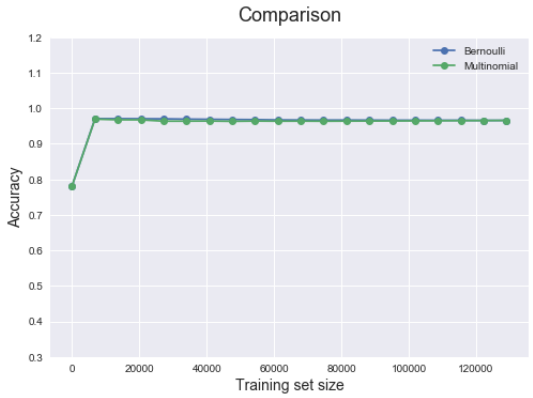

Figure 1: Comparison of both models on *Toys and Games datasets* for sentiment classification using cleaned review.

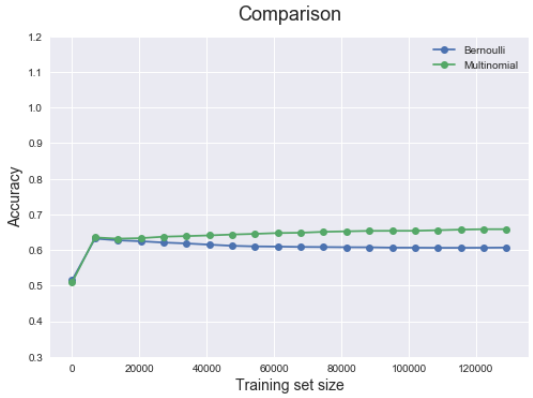

Figure 2: Comparison of both models on *Toys and Games datasets* for score classification using cleaned review.

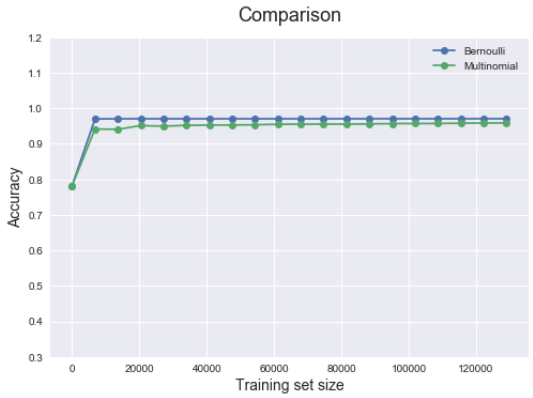

Figure 3: Comparison of both models on *Toys and Games datasets* for sentiment classification using cleaned review and only 50 words in dictionary.

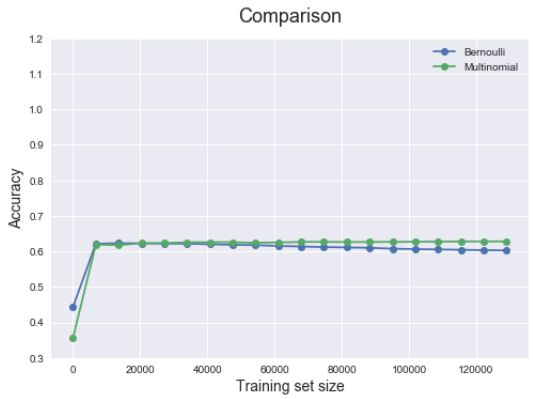

Figure 4: Comparison of both models on *Toys and Games datasets* for score classification using cleaned review and only 50 words in dictionary.

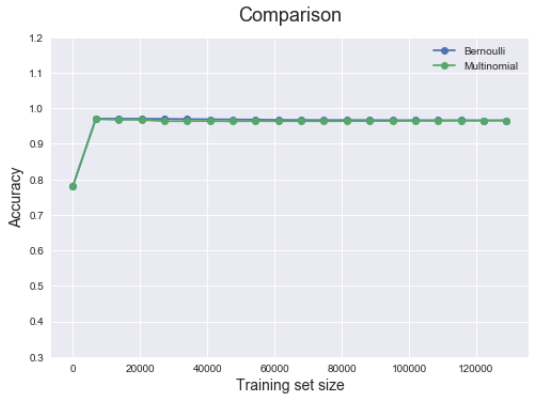

Figure 5: Comparison of both models on *Toys and Games datasets* for sentiment classification using cleaned summary reviews.

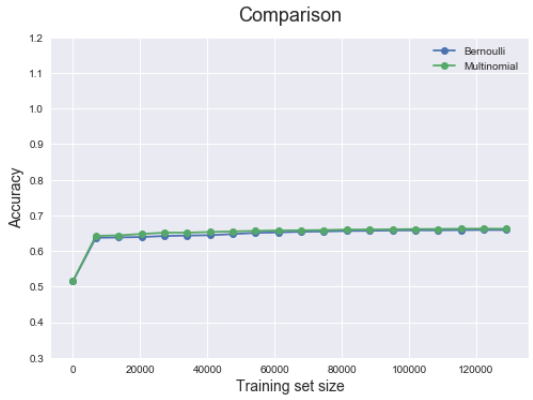

Figure 6: Comparison of both models on *Toys and Games datasets* for score classification using cleaned summary reviews.

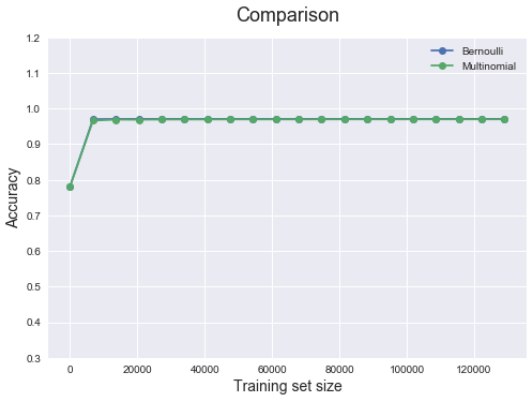

Figure 7: Comparison of both models on *Toys and Games datasets* for sentiment classification using cleaned summary reviews and only 50 words in dictionary.

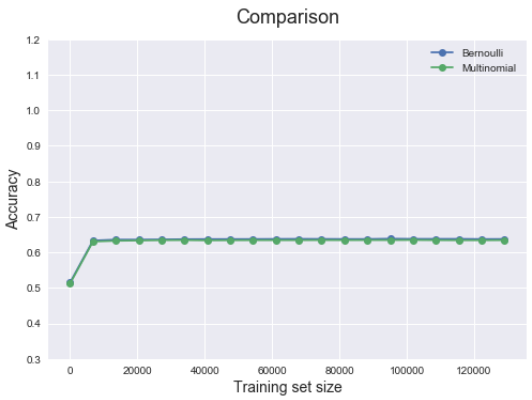

Figure 8: Comparison of both models on *Toys and Games datasets* for score classification using cleaned summary reviews and only 50 words in dictionary.

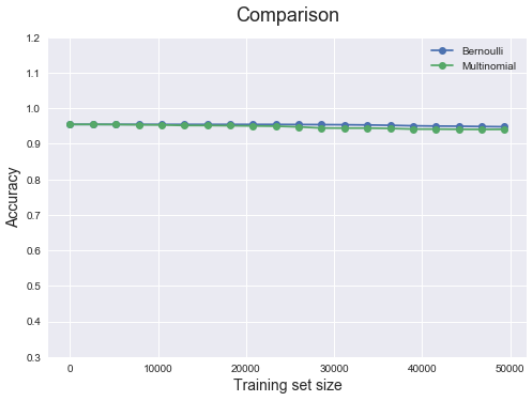

Figure 9: Comparison of both models on *Digital Music datasets* for sentiment classification using cleaned review.

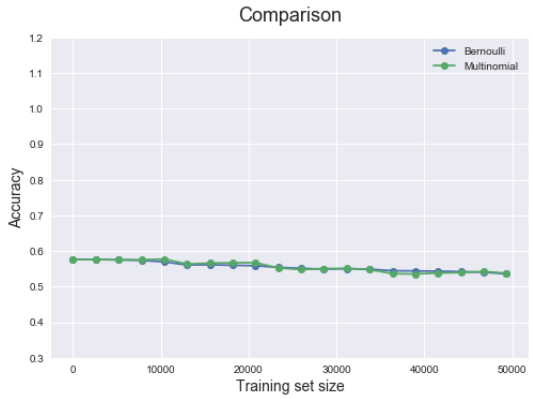

Figure 10: Comparison of both models on *Digital Music datasets* for score classification using cleaned review.

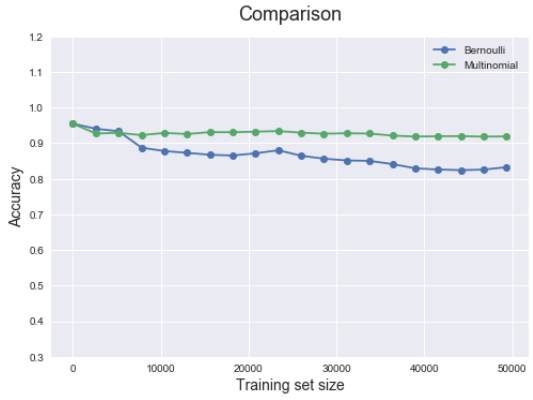

Figure 11: Comparison of both models on *Digital Music datasets* for sentiment classification using cleaned review and only 50 words in dictionary.

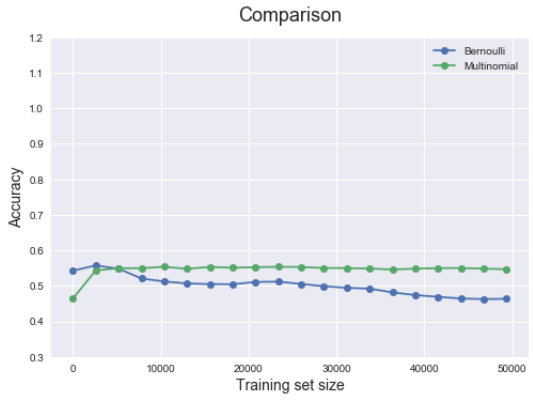

Figure 12: Comparison of both models on *Digital Music datasets* for score classification using cleaned review and only 50 words in dictionary.

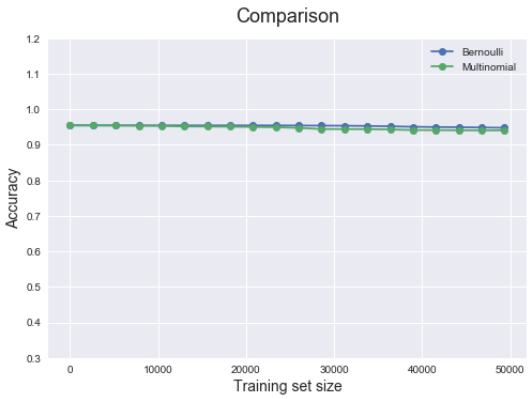

Figure 13: Comparison of both models on *Digital Music datasets* for sentiment classification using cleaned summary reviews.

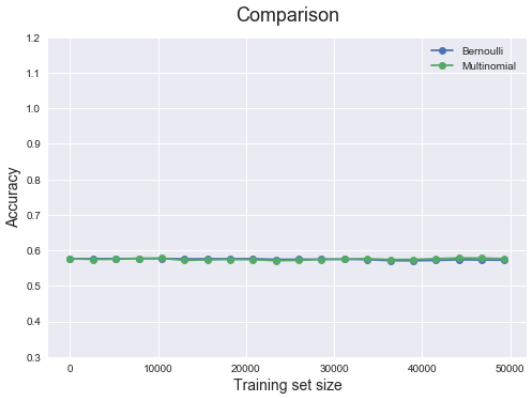

Figure 14: Comparison of both models on *Digital Music datasets* for score classification using cleaned summary reviews.

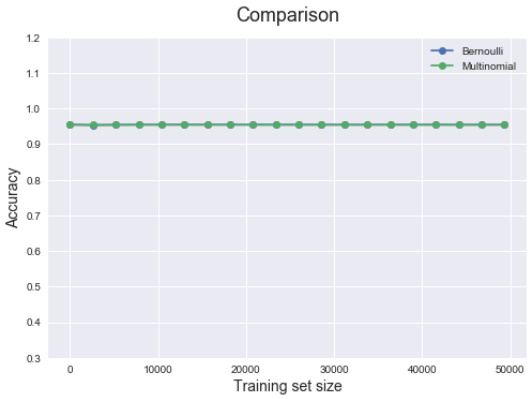

Figure 15: Comparison of both models on *Digital Music datasets* for sentiment classification using cleaned summary reviews and only 50 words in dictionary.

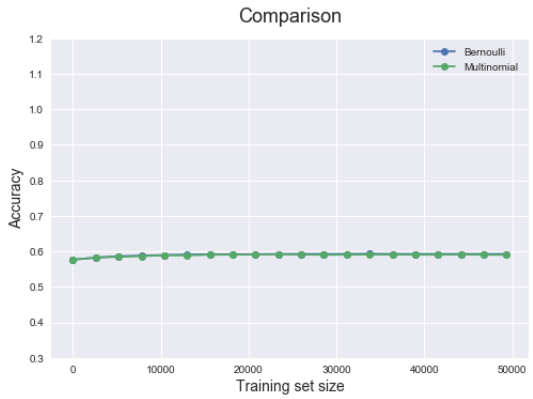

Figure 16: Comparison of both models on *Digital Music datasets* for score classification using cleaned summary reviews and only 50 words in dictionary.

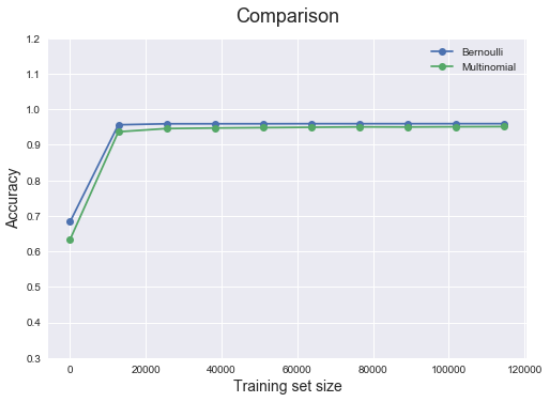

Figure 19: Comparison of both models on *Grocery and Gourmet food datataset* for sentiment classification using cleaned review and only 50 words in dictionary.

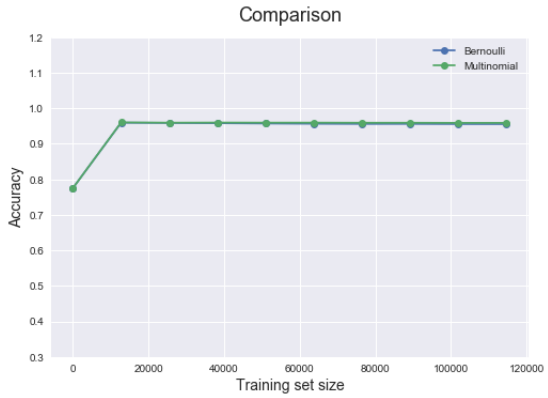

Figure 17: Comparison of both models on *Grocery and Gourmet food datataset* for sentiment classification using cleaned review.

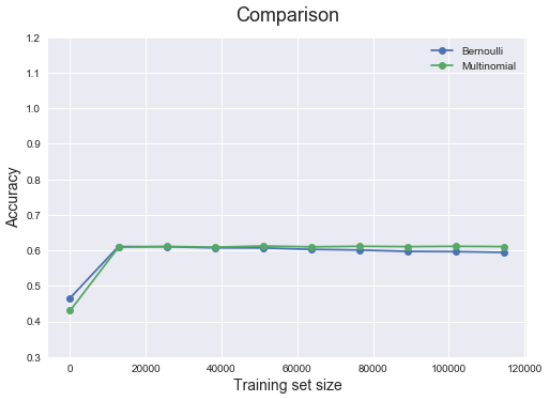

Figure 20: Comparison of both models on *Grocery and Gourmet food datataset* for score classification using cleaned review and only 50 words in dictionary.

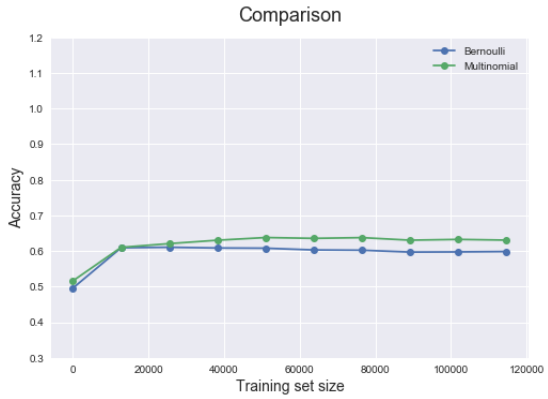

Figure 18: Comparison of both models on *Grocery and Gourmet food datataset* for score classification using cleaned review.

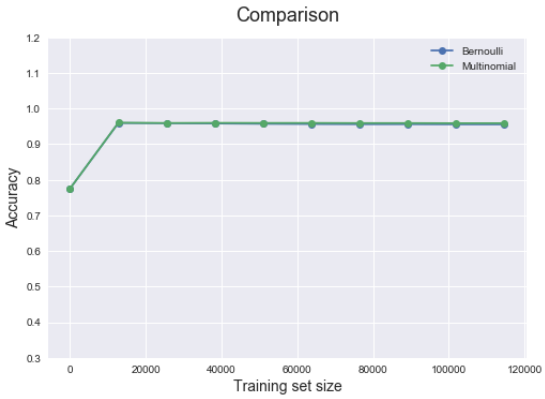

Figure 21: Comparison of both models on *Grocery and Gourmet food datataset* for sentiment classification using cleaned summary reviews.

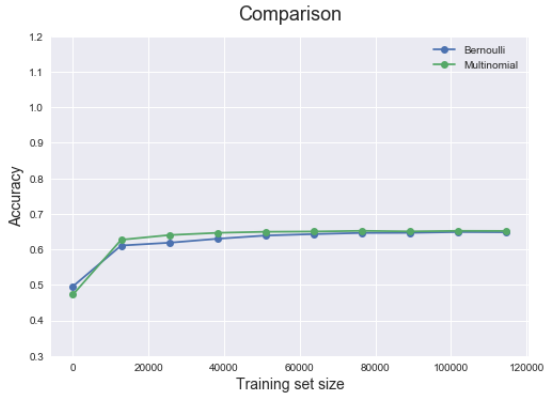

Figure 22: Comparison of both models on *Grocery and Gourmet food datataset* for score classification using cleaned summary reviews.

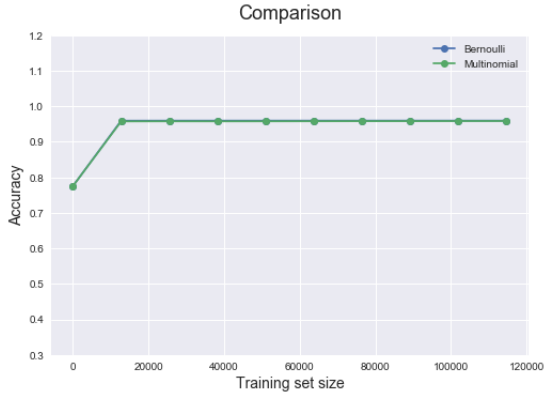

Figure 23: Comparison of both models on *Grocery and Gourmet food datataset* for sentiment classification using cleaned summary reviews and only 50 words in dictionary.

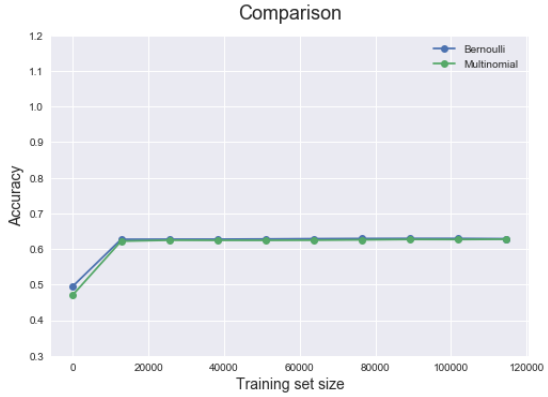

Figure 24: Comparison of both models on *Grocery and Gourmet food datataset* for score classification using cleaned summary reviews and only 50 words in dictionary.