Data:

Figure 1: Raw data

```
lie     sentiment       review
f       n       'Mike\'s Pizza High Point, NY Service was very slow and the quality was low. You would think they would know at least how to make good
pizza, not. Stick to pre-made dishes like stuffed pasta or a salad. You should consider dining else where.'
f       n       'i really like this buffet restaurant in Marshall street. they have a lot of selection of american, japanese, and chinese dishes. we also
got a free drink and free refill. there are also different kinds of dessert. the staff is very friendly. it is also quite cheap compared with the other
restaurant in syracuse area. i will definitely coming back here.'
f       n       'After I went shopping with some of my friend, we went to DODO restaurant for dinner. I found worm in one of the dishes .'
f       n       'Olive Oil Garden was very disappointing. I expect good food and good service (at least!!) when I go out to eat. The meal was cold when we
got it, and the waitor had no manners whatsoever. Don\'t go to the Olive Oil Garden. '
f       n       'The Seven Heaven restaurant was never known for a superior service but what we experienced last week was a disaster. The waiter would not
notice us until we asked him 4 times to bring us the menu. The food was not exceptional either. It took them though 2 minutes to bring us a check after they
spotted we finished eating and are not ordering more. Well, never more. '
f       n       'I went to XYZ restaurant and had a terrible experience. I had a YELP Free Appetizer coupon which could be applied upon checking in to the
restaurant. The person serving us was very rude and didn\'t acknowledge the coupon. When I asked her about it, she rudely replied back saying she had
already applied it. Then I inquired about the free salad that they serve. She rudely said that you have to order the main course to get that. Overall, I had
a bad experience as I had taken my family to that restaurant for the first time and I had high hopes from the restaurant which is, otherwise, my favorite
place to dine. '
```

Figure 2 shows a part of the raw data. Each row contains a restaurant review which is the column called review. There are 2 other columns, where only sentiment is used for the study. The sentiment column takes the value of “p” or “n”, with “p” stands for positive review, and “n” for negative review. There are 92 reviews with half positive reviews.

Before the analysis, a series of preprocessing was done on the reviews, which were converted to all lower-case letters. First there were some “\” and “’s” in the words like “he\’s” which provides no value for analysis. Hence, they were removed. Second, there were some symbols in the words likes “.”, “-”, “!”, “/”, and “=”. They were replaced with space due to the same reason as “\”. Third, stop words like “you”, “and”, etc. were removed with same reason. Forth, the words were stemmed to treat words in different forms as one. A sample of cleaned review is shown in table 1. At last, the reviews were vectorized.

There are 2 methods used for vectorization in the study. First the reviews were tokenized, and each got counted by occurrence in each review. Second the reviews were tokenized, and each token gets a value of 1 if it exists in a review, or 0 if not exists. Table 2 and 3 shows how the vectorized reviews look like with each row represents a review.

Table 1: Cleaned reviews

sentiment	review
n	Mike high point slow quality low would think w...

Table 2:Count vectorization

clean	good	delicious	dirty	the	a
0	0	0	0	1	1
2	0	0	4	0	2
0	0	1	1	1	5

Table 3: Boolean vectorization

clean	good	delicious	dirty	the	a
0	0	0	0	1	1
1	0	0	1	0	1
0	0	1	1	1	1

Figure 2:Positive reviews

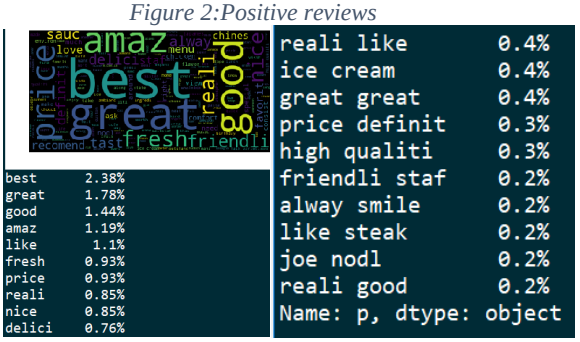

Figure 3: Negative reviews

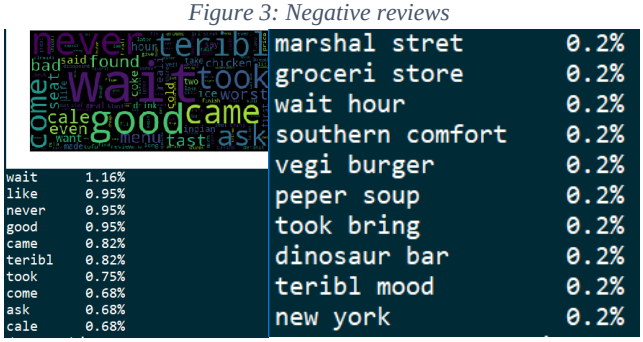

Figure 6 and 7 show the word cloud and word frequencies for the positive and negative reviews. They seem to be very different. The unigram words seem to be very informative for both positive and negative reviews. This could mean that unigrams can be good predictors for the classification model. On the other hand, looking at top bigrams for positive reviews, they look very informative. However, the frequencies are very low, and there's not much variation. Looking at bigrams for negative reviews, they don't seem informative. The frequencies seem to be having no variation. This is a sign that bigrams might not be good predictors for sentiment analysis.

Model:

The model used in the study is the Multinomial Naïve Bayes Classifier. This is a supervised machine learning method, where the goal is to predict whether a review is positive or negative and lie or truth. Its algorithm is based on the Statistics theorem called Baye's Theorem. This theorem allows us to calculate conditional probability of "A" given "B" is true(P(A|B)). The formula is as fallow: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. In words, it means probability of "A" given "B" can be calculated using probability of "B" given "A", "A", and "B". In this case P(A|B) is called posterior probability, and the rest are called prior probability. In the model, "A" can be thought of as a label, and "B" as a review. However, in the study the reviews were break down into tokens instead of sentences. Now P(B|A) becomes $P(B_0, B_1, B_2, \dots B_n|A)$. In Statistics, $P(B_0, B_1, B_2, \dots B_n|A) = \prod_{i=0}^{i=n} P(B_i|A)$ if all the "B"s are independent. However, in any sentence the words are not independent. Fortunately, based on some studies, it was found in text mining even if the assumption of independence is violated, M.N.B can still yield good results. As another note, $P(B_1|A)$ can be thought of as probability of the word B_1 given the label A. Assuming there are 100 words in reviews labeled A, and the word B_1 occurred 10 times out of these reviews, then $P(B_1|A) = 0.1$. However, $P(B_1)$ can't be calculated. Thankfully for task of classification, the need is to predict the label. All $P(B|A)P(A)$ can be calculated, so to predict the label "A", simply pick the maximum value of $P(B|A)P(A)$.

10 different Naïve Bayes models were built in this study for comparison. The details of the models are shown in table 4. The method column shows whether unigram(single words as tokens) or bigrams(single and double words as tokens). The cleaning method column shows whether a manual cleaning was performed or cleaning by Sklearn package was performed. The Multinomial Naive Bayes model takes word counts as inputs, while Bernoulli Naive Bayes

model takes boolean expressions (0 means not exist; 1 means exists) as input.

Table 4:Models details

Model	Method	Cleaning Method	Remove stop words
Multinomial Naive Bayes	Unigram	Sklearn	Y
Multinomial Naive Bayes	Unigram	Sklearn	N
Multinomial Naive Bayes	Unigram	Manual	N
Multinomial Naive Bayes	Unigram	Manual	Y
Multinomial Naive Bayes	Bigrams	Sklearn	Y
Multinomial Naive Bayes	Bigrams	Manual	Y
Bernoulli Naive Bayes	Unigram	Manual	Y
Bernoulli Naive Bayes	Unigram	Sklearn	Y
Bernoulli Naive Bayes	Bigrams	Sklearn	Y
Bernoulli Naive Bayes	Bigrams	Manual	Y

The evaluation method used in this study is the hold out test accuracy score. This study split the data into 80-20% for training and testing.

Results:

Table 5: Models results

Model	Method	Cleaning Method	Remove stop words	Test Score
Multinomial Naive Bayes	Unigram	Sklearn	Y	0.894737
Multinomial Naive Bayes	Unigram	Sklearn	N	0.894737
Multinomial Naive Bayes	Unigram	Manual	N	0.894737
Multinomial Naive Bayes	Unigram	Manual	Y	0.894737
Multinomial Naive Bayes	Bigrams	Sklearn	Y	0.842105
Multinomial Naive Bayes	Bigrams	Manual	Y	0.842105
Bernoulli Naive Bayes	Unigram	Manual	Y	0.842105
Bernoulli Naive Bayes	Unigram	Sklearn	Y	0.789474
Bernoulli Naive Bayes	Bigrams	Sklearn	Y	0.526316
Bernoulli Naive Bayes	Bigrams	Manual	Y	0.526316

Figure 4: Models Results

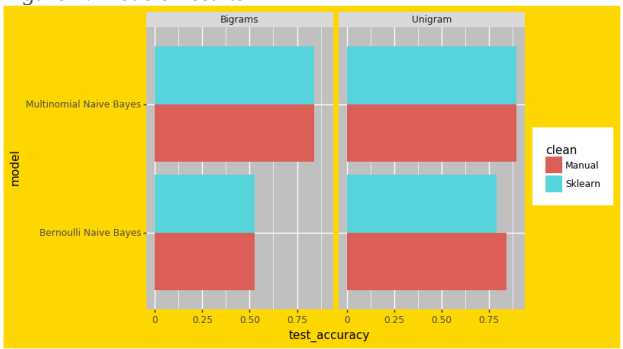

Figure 5: Models Results

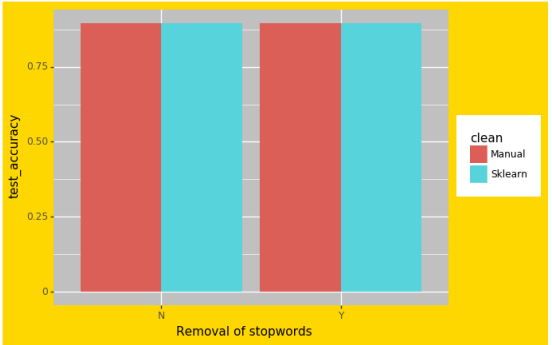

Table 5, figure 5, and figure 6 show the results of different Naïve Bayes models. It's clear that overall Multinomial Naïve Bayes performs better than Bernoulli Naive Bayes. It's also

showing that by using unigram, the testing accuracy is higher than including bigrams. This could be due to the size of the text for each data point. Since the analysis for this study is done on restaurant reviews, the texts are not likely to be long. In short texts, there are just no enough words to get enough bigrams for analysis purpose. In long text documents, there could be enough bigrams to be extracted with high variation of bigrams frequencies. Next looking at the cleaning method, in general there don't seem to be any difference between manual cleaning and using Sklearn package's function. However, for the unigram model with Bernoulli Naive Bayes, manual cleaning seems to be performing better than using Sklearn package. Lastly looking at whether to remove stop words or not, it seems like whether to remove stop words don't make any difference. This again could be due to the text size. The stop words size would increase as the text size, and perhaps it would start having effects on the models.

Overall Multinomial Naïve Bayes using unigram with Sklearn without removing stop words might be the best way to go for sentiment analysis. This is assuming only Naïve Bayes models being considered. As a note, according to Occam's Razor law, "Among competing hypotheses that predict equally well, the one with the fewest assumptions should be selected.". In the case of model selection in machine learning, if there are several models performing equally well, pick the simplest one.