**Project 2**

**Naive Bayes algorithm for learning to classify text**

Latheesh Mangeri

1002116009

**Introduction:**

Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents. The dataset containing 20,000 newsgroup messages drawn from the 20 newsgroups.

**Dataset:**

The dataset contains 1000 documents from each of the 20 newsgroups such as ['talk.politics.mideast', 'rec.autos', 'comp.sys.mac.hardware', 'alt.atheism', 'rec.sport.baseball', 'comp.os.ms-windows.misc', 'rec.sport.hockey', 'sci.crypt', 'sci.med', 'talk.politics.misc', 'rec.motorcycles', 'comp.windows.x', 'comp.graphics', 'comp.sys.ibm.pc.hardware', 'sci.electronics', 'talk.politics.guns', 'sci.space', 'soc.religion.christian', 'misc.forsale', 'talk.religion.misc']

**Implementation:**

**Data Importing:**

The data from dataset will be imported in a dictionary. Dictionary keys have 20 newsgroups. Dictionary values have the text that has been taken from each file.

**Data Preprocessing:**

The text taken from each file will undergo preprocessing using

* Lower(Lowering the Text)
* Special Characters(Removing Special Characters)
* Tokenize
* Stopwords(Removing Extended Stopwords)
* Lemmatization

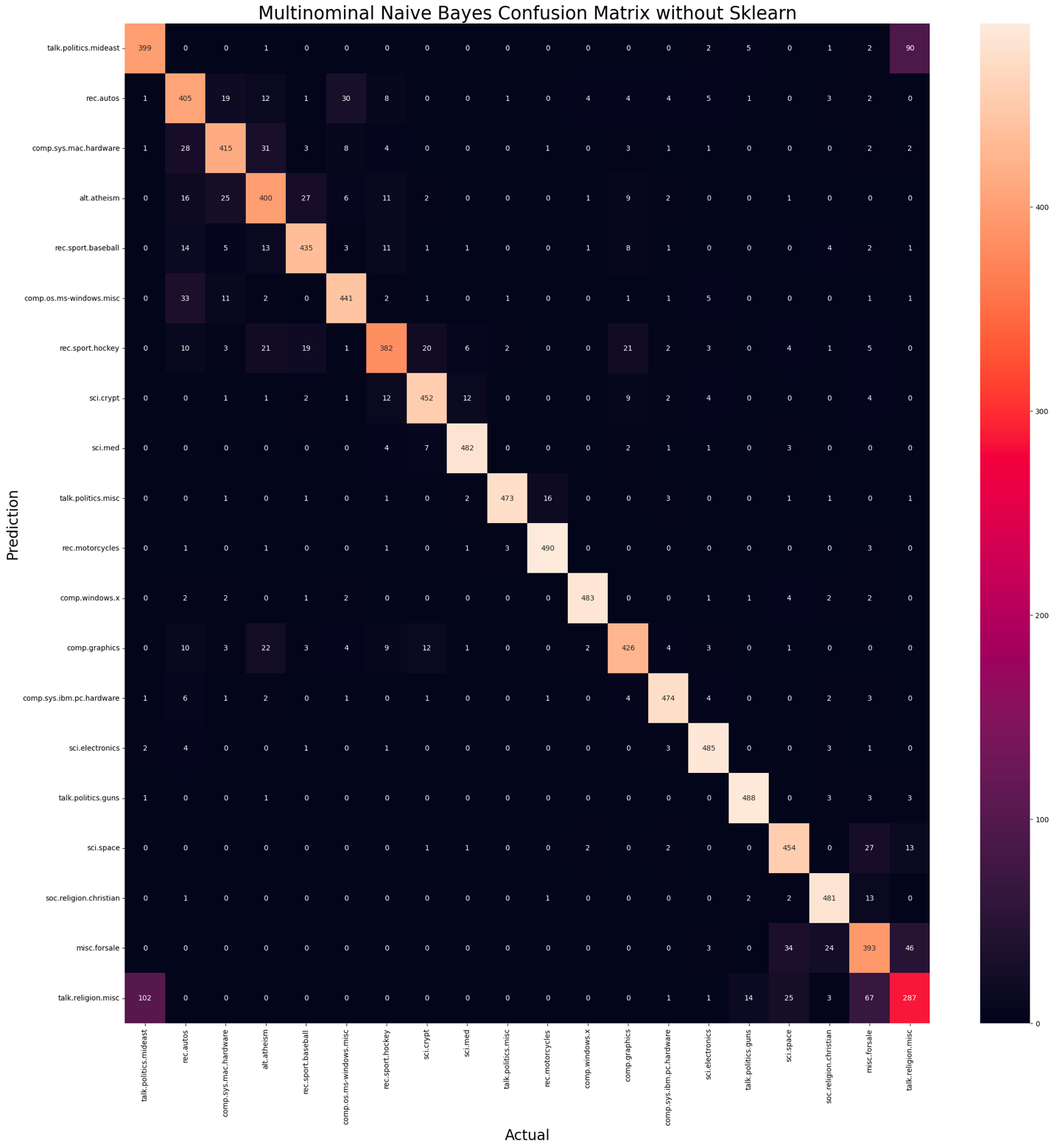
**Data Splitting:**

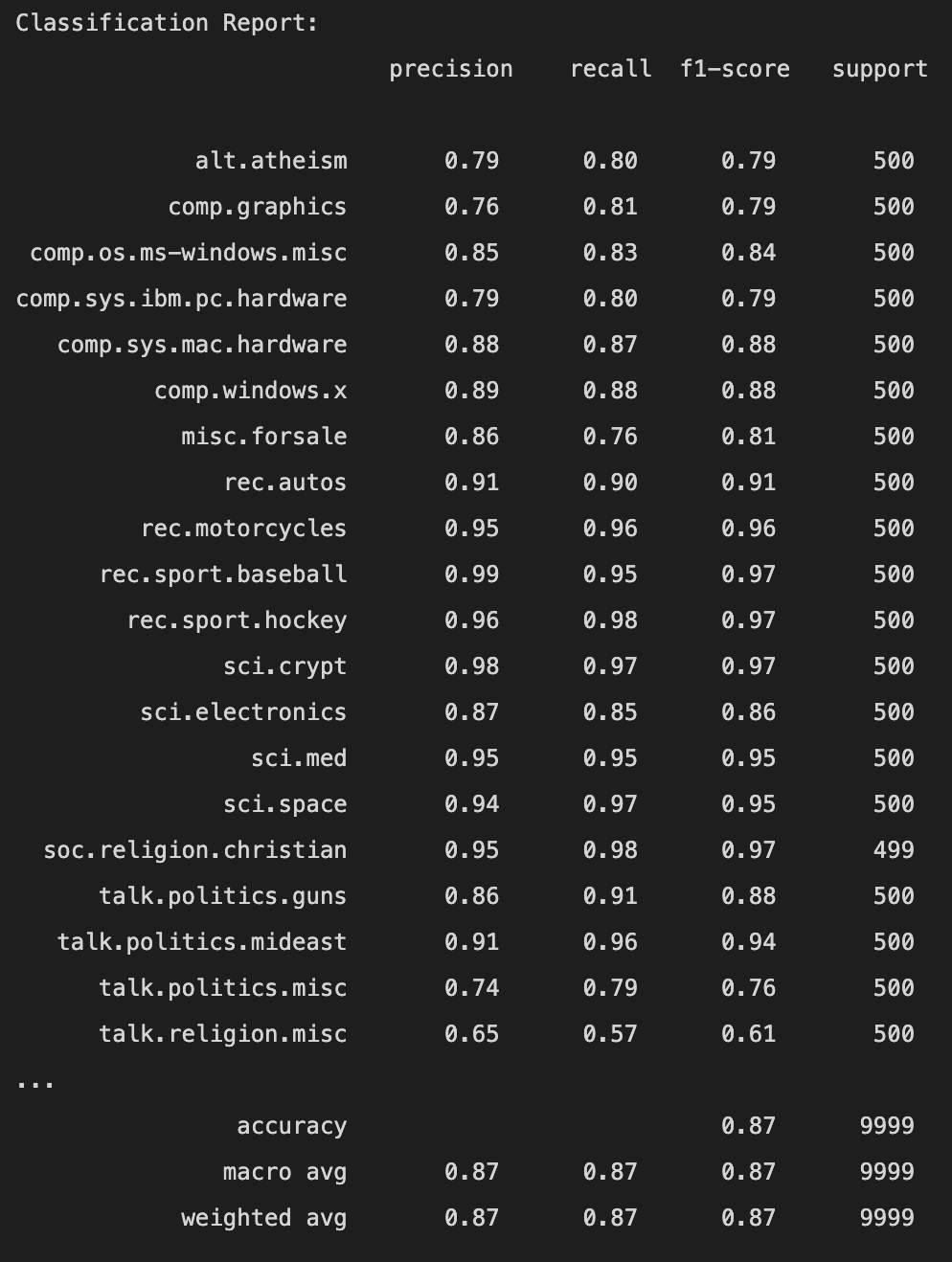
The dictionary preprocessed data will be split as 50% train data, 50% test data. The split will be done by taking 50% of each group in each file to train and rest 50% will be under test data.

**Multinominal Naïve Bayes:**

I have implemented multinominal Naïve Bayes algorithm using Sklearn and also without Sklearn. I am attaching the results below.

**Confusion Matrix without Sklearn:**



**Classification Report: **

**Conclusion:**

|  |  |  |
| --- | --- | --- |
|  | Multinominal NB(using Sklearn) | Multinominal NB(without Sklearn) |
| Test Accuracy | 0.90279 | 0.874587 |
| Train Accuracy | 0.963293 | 0.986397 |