# Text Classification with Naive Bayes

**20221009 GKS Kim Eunhong**

**1. Introduction**

Text classification is a natural language processing (NLP) task that involves assigning predefined categories or labels to text documents. In this implementation, I utilize the Naive Bayes classifier, a probabilistic algorithm based on Bayes' theorem, which assumes that the features (words in this case) are conditionally independent. The Naive Bayes algorithm is commonly used for text classification tasks due to its simplicity and efficiency. This text classification systems also lies the Naive Bayes algorithm, a probabilistic model rooted in Bayes' theorem.

Bayesian Foundation: Bayesian methods are grounded in probability theory, particularly Bayes' theorem, which calculates the probability of a hypothesis given observed evidence. In the context of text classification, the hypothesis is the category to which a document belongs, and the evidence is the content of the document itself.

Naive Bayes Assumption: The "naive" in Naive Bayes arises from the assumption of conditional independence among features, in this case, the words or tokens in a document. Despite this simplification, Naive Bayes has proven remarkably effective for text classification tasks, demonstrating its resilience and efficiency in the face of large and diverse datasets.

**2. Implementation:**

2.1 Data Fetching:

I used the `fetch_20newsgroups` dataset from scikit-learn, which contains a collection of approximately 20,000 newsgroup documents across 20 different categories.

2.2 Vocabulary Creation:

I constructed a vocabulary set by tokenizing and lowercasing words from the training dataset. The unique words in the vocabulary set were used to create dictionaries for storing word counts and log probabilities.

2.3 Model Training:

For each category, Laplace smoothing was applied to the word counts, and the probabilities were computed. Log probabilities were then calculated for each word, providing a more stable measure for very small probabilities.

2.4 Model Testing:

The trained model was applied to the test dataset. For each test sample, log probabilities for each category were computed, and the category with the maximum log probability was selected as the predicted category.

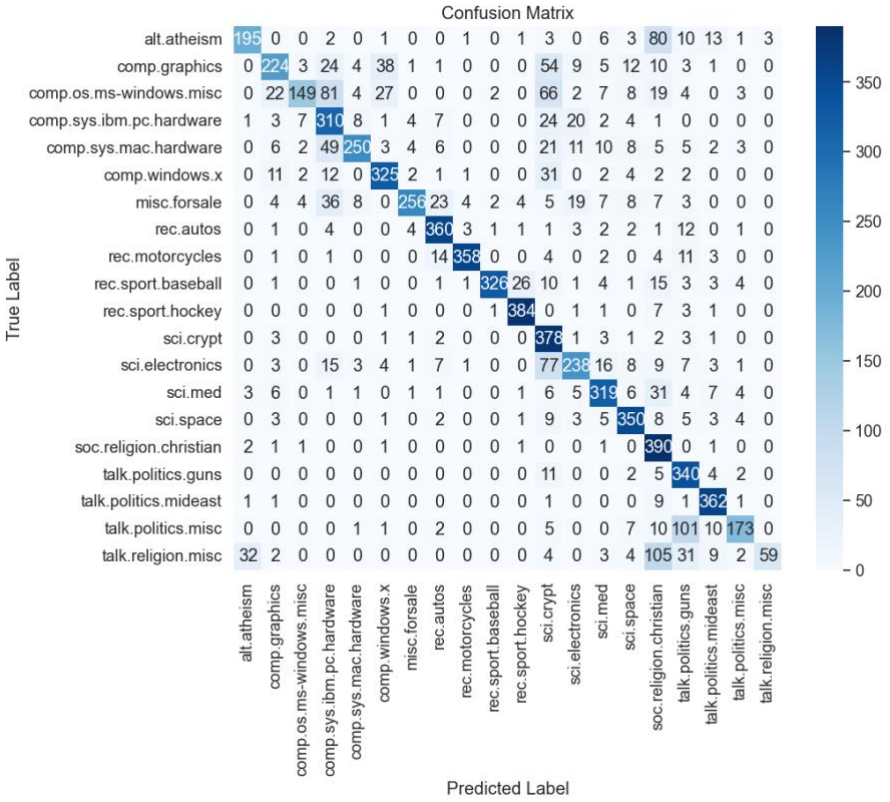
**3. Performance Evaluation**

3.1 Accuracy:

The accuracy of the model on the test dataset was calculated and printed. The accuracy is a measure of the proportion of correctly predicted labels among all test samples. The measured model accuracy was about 76%.

3.2 Confusion Matrix Visualization:

A confusion matrix was generated to provide a detailed breakdown of correct and incorrect predictions across different categories. The confusion matrix was visualized as a heatmap using seaborn.


Confusion Matrix
True Label
Predicted Label
alt.atheism 195 0 0 2 0 1 0 0 1 0 1 3 0 6 3 80 10 13 1 3
comp.graphics 0 224 3 24 4 38 1 1 0 0 0 54 9 5 12 10 3 1 0 0
comp.os.ms-windows.misc 0 22 149 81 4 27 0 0 0 2 0 66 2 7 8 19 4 0 3 0
comp.sys.ibm.pc.hardware 1 3 7 310 8 1 4 7 0 0 0 24 20 2 4 1 0 0 0 0
comp.sys.mac.hardware 0 6 2 49 250 3 4 6 0 0 0 21 11 10 8 5 5 2 3 0
comp.windows.x 0 11 2 12 0 325 2 1 1 0 0 31 0 2 4 2 2 0 0 0
misc.forsale 0 4 4 36 8 0 256 23 4 2 4 5 19 7 8 7 3 0 0 0
rec.autos 0 1 0 4 0 0 4 360 3 1 1 1 3 2 2 1 12 0 1 0
rec.motorcycles 0 1 0 1 0 0 0 14 358 0 0 4 0 2 0 4 11 3 0 0
rec.sport.baseball 0 1 0 0 1 0 0 1 1 326 26 10 1 4 1 15 3 3 4 0
rec.sport.hockey 0 0 0 0 0 1 0 0 0 1 384 0 1 1 0 7 3 1 0 0
sci.crypt 0 3 0 0 0 1 1 2 0 0 0 378 1 3 1 2 3 1 0 0
sci.electronics 0 3 0 15 3 4 1 7 1 0 0 77 238 16 8 9 7 3 1 0
sci.med 3 6 0 1 1 0 1 1 0 0 1 6 5 319 6 31 4 7 4 0
sci.space 0 3 0 0 0 1 0 2 0 0 1 9 3 5 350 8 5 3 4 0
soc.religion.christian 2 1 1 0 0 1 0 0 0 0 1 0 0 1 0 390 0 1 0 0
talk.politics.guns 0 0 0 0 0 0 0 0 0 0 0 11 0 0 2 5 340 4 2 0
talk.politics.mideast 1 1 0 0 0 0 0 0 0 0 0 1 0 0 0 9 1 362 1 0
talk.politics.misc 0 0 0 0 1 1 0 2 0 0 0 5 0 0 7 10 101 10 173 0
talk.religion.misc 32 2 0 0 0 0 0 0 0 0 0 4 0 3 4 105 31 9 2 59
alt.atheism
comp.graphics
comp.os.ms-windows.misc
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
comp.windows.x
misc.forsale
rec.autos
rec.motorcycles
rec.sport.baseball
rec.sport.hockey
sci.crypt
sci.electronics
sci.med
sci.space
soc.religion.christian
talk.politics.guns
talk.politics.mideast
talk.politics.misc
talk.religion.misc
350
300
250
200
150
100
50
0