# Naive Bayes Classifier

DSCI-575
Spring 2021

Hayden Sather

## I. Introduction

Naive Bayes is a very simple, yet effective classification algorithm. It is especially useful for classifying text data into discrete categories. This project, the Naive Bayes classifier was used to classify a dataset of over 18,000 messages sent from 20 different news groups (scikit-learn, 2017). The repository can be found online at the first link in the Bibliography (Sather, 2021).

## II. Training

### A. Data

The Naive Bayes classifier was trained on 60% percent of the data which corresponds to 11314 messages. This leaves 7532 messages for testing.

### B. Common Word Exclusion

Additionally, text files that include 1025 of the most common English words, 1000 of the most common male names (namecensus, 2020), 1000 of the most common female names, and the 1000 most common last names (namecensus, 2016) were used to limit common words that were used in the classification process. Only words that were not in these common lists were included and considered in the classification algorithm.

## III. Evaluation

### A. One-to-Many Classification

The first test that was ran was the One-to-Many Classification. All of the testing messages and all of the classes were considered. Each message was then assigned to a class by the prediction algorithm. The results are summarized below. As can be seen, these accuracies are very high and are roughly normal and symmetrical. This is an indication that in a many to one classification setting, Naive Bayes works

*1) Overall Accuracy:* The overall accuracy of the Naive Bayes Classifier in a One-to-Many problem was 97%.

*2) Individual Accuracy:* The individual accuracy scores can be shown below in Figure 1.

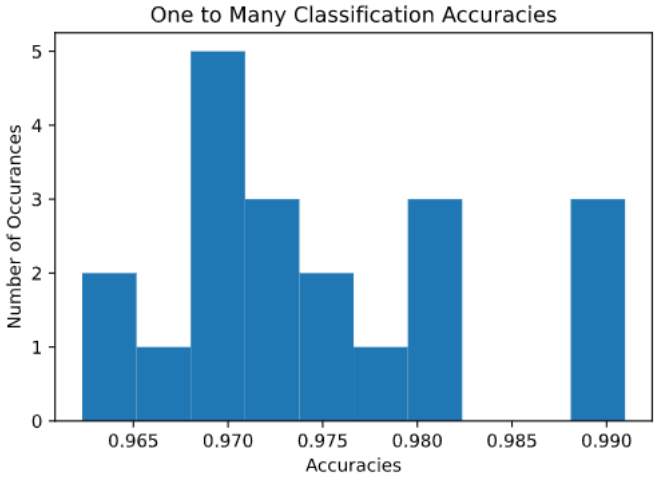


Fig. 1. Histogram of Accuracies from One-to-Many Classifier

### B. Binary Classification

The next test was to perform pairwise binary classification. Each combination of pairs of classifiers was considered for this problem. The prediction algorithm picked which of the two classes was the most likely. Only the messages that were between the two were considered in testing this classifier. The results are shown below.

*1) Histogram of Accuracies:* The histogram of accuracies can be seen in Figure 2. This is roughly bimodal, as the lower accuracies are due to classes that are very similar, as will be shown next.

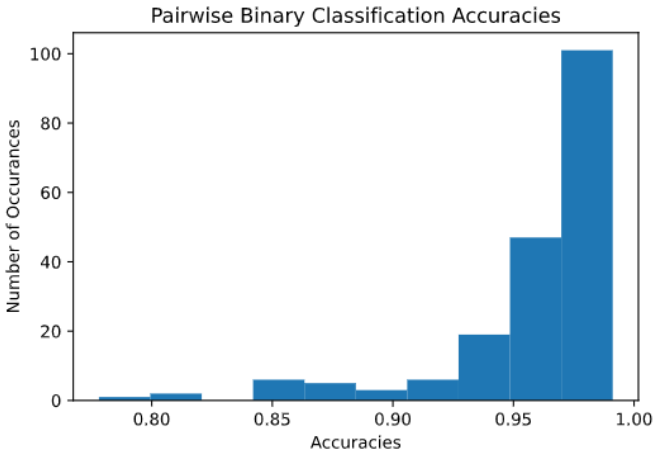


Fig. 2. Histogram of Accuracies from Bimodal Classifier

*2) Highest and Lowest Accuracies:* The pairs of classes with the top 5 accuracies can be seen in Table I in descending order. As expected, these classes are very different from each other and have very little in common.

The pairs of classes with the lowest 5 accuracies can be seen in Table II in descending order. As expected, these classes are

| |
|---|
| comp.sys.ibm.pc.hardware , rec.sport.hockey |
| alt.atheist , comp.sys.ibm.pc.hardware |
| rec.sport.hockey , sci.space |
| comp.os.ms-windows.misc , rec.sport.hockey |
| comp.os.ms-windows.misc , rec.motorcycles |

TABLE I
THE 5 PAIRS WITH BEST ACCURACY

very similar to each other and have a lot in common.

| |
|---|
| soc.religion.christian , talk.religion.misc |
| talk.politics.guns , talk.politics.misc |
| alt.atheism , talk.religion.misc |
| talk.politics.misc , talk.religion.misc |
| talk.politics.mideast , talk.politics.misc |

TABLE II
THE 5 PAIRS WITH WORST ACCURACY

### C. Trinary Classification

The next test was to perform trinary classification. Each combination of groups of 3 classifiers was considered for this problem. The prediction algorithm picked which of the three classes was the most likely. Only the messages that were between the three were considered in testing this classifier. The results are shown below.

*1) Histogram of Accuracies:* The histogram of accuracies can be seen in Figure 3. This histogram is much more normally distributed, albeit skewed left. This is likely representative of the fact that the comparison of groups of classes is more likely to return roughly similar results because it is much more difficult to have 3 classes that are very different or very similar than it is to have 2 classes that are very different or very similar.

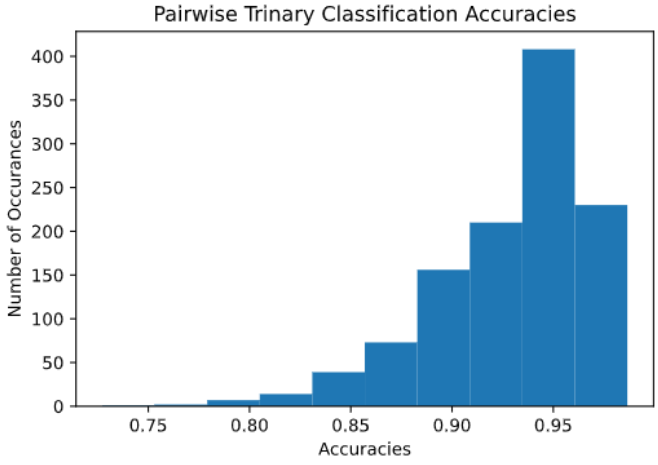


Fig. 3. Histogram of Accuracies from Trimodal Classifier

*2) Highest and Lowest Accuracies:* The groups of classes with the top 5 accuracies can be seen in Figure III. As expected, these classes are very different from each other and have very little in common.

| |
|---|
| comp.os.ms-windows.misc , rec.motorcycles , rec.sport.hockey |
| comp.sys.ibm.pc.hardware , rec.motorcycles , rec.sport.hockey |
| comp.windows.x , rec.motorcycles , rec.sport.hockey |
| rec.motorcycles , rec.sport.hockey , sci.space |
| comp.sys.mac.hardware , rec.motorcycles , rec.sport.hockey |

TABLE III
THE 5 GROUPS WITH BEST ACCURACY

The groups of classes with the lowest 5 accuracies can be seen in Figure IV. As expected, these classes are very similar to each other and have a lot in common.

| |
|---|
| alt.atheism , soc.religion.christian , talk.religion.misc |
| talk.politics.guns , talk.politics.misc , talk.religion.misc |
| alt.atheism , talk.politics.misc , talk.religion.misc |
| talk.politics.mideast , talk.politics.misc , talk.religion.misc |
| alt.atheism , talk.politics.mideast , talk.religion.misc |

TABLE IV
THE 5 GROUPS WITH WORST ACCURACY

## IV. DISCUSSION

### A. Bi vs. Tri Models

The biggest point to take away from this experiment is that trinary classification provides results with less variance and a decrease in accuracy than binary classification does. With binary classification, it is easier for two samples to be very similar or very different from each other. It is less likely that three samples are very similar or very different, especially because text is such a high dimensional piece of input data. As the number of members of a group rises, the more the distribution of classification accuracies varies, slightly decreases, and becomes non-normal

### B. Accuracy Explanations

The pairs of data that had the worst accuracy were those that were similar to each other. The closer that the messages are, in terms of word probability space, the harder it was for the Naive Bayes algorithm to decide on an effective decision border. One example is that hockey and PC hardware have the highest accuracy when deciding between them, and that they are both very unrelated. On the other hand, the more that two messages do not share the same word probability make up, the easier it is for the Naive Bayes algorithm to decide on an effective decision border. An example of this is that the guns and the politics class had the best accuracy, and guns and politics are intrinsically linked. The tri-class models had a similar behavior. When classes were all related, such as atheism, christian, and religion, there was a low classification accuracy because these classes had a lot of the same words. On the other hand, when the classes were not related, such as ms-windows, motorcycles, and hockey, it had a low classification accuracy because the words were too close in higher dimensional space, which made it hard to draw a decision boundary for.

## REFERENCES


[1] Sather, Hayden. "Hrsather/Naive_Bayes_Newsgroup." GitHub, 10 May 2021, github.com/hrsather/Naive_Bayes_Newsgroup.

[2] "5.6.2. The 20 Newsgroups Text Dataset." Scikit, 2017, scikit-learn.org/0.19/datasets/twenty_newsgroups.html.

[3] "Most Common Male First Names in the United States." Namecensus.com, Namecensus, 15 Apr. 2020, namecensus.com/male_names.htm.

[4] "Most Common Female First Names in the United States." Namecensus.com, Namecensus, 15 Apr. 2020, namecensus.com/female_names.htm.

[5] Namecensus. "Most Common Last Names for People of Two or More Races in the United States." Namecensus.com, Namecensus, 10 Mar. 2016, namecensus.com/data/two_race.html.