Assessed Lab 1

Michael McMillan - 8116775

Introduction to the Lab and Dataset

This lab focuses on text classification using the Naïve Bayes Classifier, a fundamental Natural Language Processing (NLP) task. The objective is categorising text documents into predefined classes, a process widely used in business intelligence and sentiment analysis.

We use the 20 Newsgroups dataset, which contains around 20,000 documents across 20 topics, as a benchmark for our classification task. The lab involves implementing and evaluating a Multinomial Naïve Bayes classifier with a Bag of Words model and comparing its performance with a Complement Naïve Bayes classifier.

1. **Describe the critical steps for data preparation and feature extraction (1 mark).**

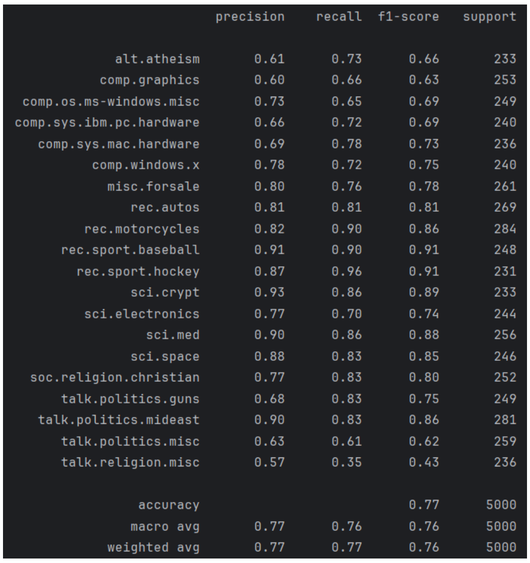
Preparing Data

- **Data Collection:** Gather the dataset and ensure it is complete and representative of the issue area. For instance, downloading tagged photos from a dataset like Kaggle could be necessary for image classification.
- **Data cleaning** includes filling in the blanks, eliminating duplicates, and fixing mistakes. It could also entail normalising mismatched formats, removing outliers, and filling in missing numbers.
- **Data Splitting:** Create test, validation, and training sets from the dataset. This makes it easier to assess the model's performance objectively.
- **Data Augmentation:** (Relating to picture data). For this stage, we must use modifications like rotation, flipping, scaling, and cropping to expand the dataset and provide additional unpredictability.

Feature Extraction

- **Feature Selection:** Determine which pertinent features impact the target variable most and pick those. Depending on subject expertise, this may entail manual selection or automated techniques like Principal Component Analysis (PCA).
- Feature Engineering: To enhance model performance, create new features from preexisting ones. For text categorisation, for instance, you may use methods such as TF-IDF to translate textual data into numerical features.
- Normalisation/Standardisation: Scale attributes to a comparable range, particularly for algorithms (like kNN) that depend on distance measures. Usually, this entails unit variance and a mean of zero for the features.
- Dimensionality Reduction: To lessen the impact of dimensionality, decrease the quantity of characteristics. While maintaining important information, pCA and t-SNE (t-distributed Stochastic Neighbour Embedding) can project high-dimensional data into smaller dimensions.

2. Report the overall classification results, including precision, recall, and f1-score. Explain the meaning of these criteria (1 mark).



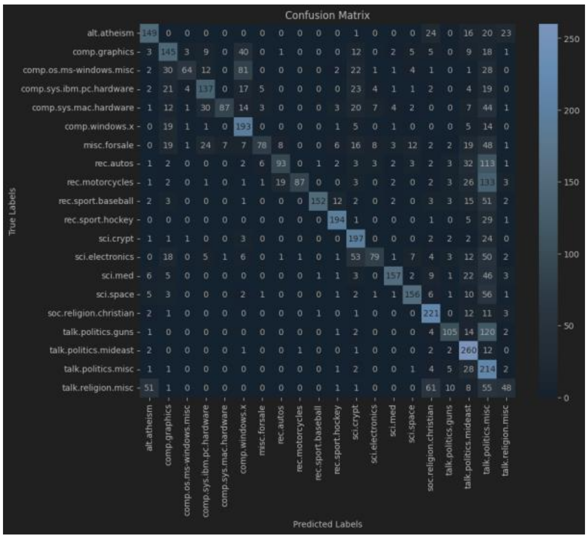
```
                          precision    recall  f1-score   support

             alt.atheism       0.61      0.73      0.66       233
           comp.graphics       0.60      0.66      0.63       253
 comp.os.ms-windows.misc       0.73      0.65      0.69       249
comp.sys.ibm.pc.hardware       0.66      0.72      0.69       240
   comp.sys.mac.hardware       0.69      0.78      0.73       236
          comp.windows.x       0.78      0.72      0.75       240
            misc.forsale       0.80      0.76      0.78       261
               rec.autos       0.81      0.81      0.81       269
         rec.motorcycles       0.82      0.90      0.86       284
      rec.sport.baseball       0.91      0.90      0.91       248
        rec.sport.hockey       0.87      0.96      0.91       231
               sci.crypt       0.93      0.86      0.89       233
         sci.electronics       0.77      0.70      0.74       244
                 sci.med       0.90      0.86      0.88       256
               sci.space       0.88      0.83      0.85       246
  soc.religion.christian       0.77      0.83      0.80       252
      talk.politics.guns       0.68      0.83      0.75       249
   talk.politics.mideast       0.90      0.83      0.86       281
      talk.politics.misc       0.63      0.61      0.62       259
      talk.religion.misc       0.57      0.35      0.43       236

                accuracy                           0.77      5000
               macro avg       0.77      0.76      0.76      5000
            weighted avg       0.77      0.77      0.76      5000
```

Explanation of precision, recall, f1-score and support

- **Precision:** The ratio of trustworthy, accurate positive predictions to the total number of optimistic predictions (true positives + false positives). It then measures the accuracy of the optimistic predictions.
- **Recall:** The ratio of accurate optimistic predictions to the total number of actual positives (true positives + false negatives). It measures the ability to capture all relevant instances.
- **F1-Score:** The harmonic mean of precision and recall, balancing them. It is useful when the class distribution is uneven.
- **Support:** The number of actual class occurrences in the specified dataset.

These metrics help evaluate a classification model's performance, providing insights into its accuracy and reliability across different categories.

3. **<u>Plot the confusion matrix for your classification result. Find the pair of classes that confuses the classifier the most. Is this result consistent with your expectations? (1 mark).</u>**

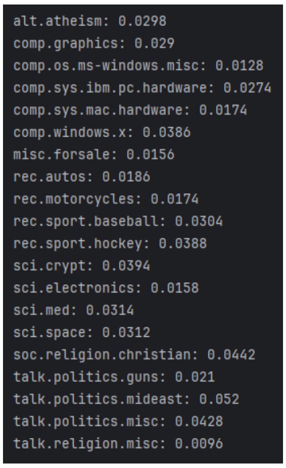


	alt.atheism	comp.graphics	comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware	comp.windows.x	misc.forsale	rec.autos	rec.motorcycles	rec.sport.baseball	rec.sport.hockey	sci.crypt	sci.electronics	sci.med	sci.space	soc.religion.christian	talk.politics.guns	talk.politics.mideast	talk.politics.misc	talk.religion.misc
alt.atheism	149	0	0	0	0	0	0	0	0	0	0	1	0	0	0	24	0	16	20	23
comp.graphics	3	145	3	9	0	40	0	1	0	0	0	12	0	2	5	5	0	9	18	1
comp.os.ms-windows.misc	2	30	64	12	0	81	0	0	0	0	2	22	1	1	4	1	0	1	28	0
comp.sys.ibm.pc.hardware	2	21	4	137	0	17	5	0	0	0	0	23	4	1	1	2	0	4	19	0
comp.sys.mac.hardware	1	12	1	30	87	14	3	0	0	0	3	20	7	4	2	0	0	7	44	1
comp.windows.x	0	19	1	1	0	193	0	0	0	0	1	5	0	1	0	0	0	5	14	0
misc.forsale	0	19	1	24	7	7	78	8	0	0	6	16	8	3	12	2	2	19	48	1
rec.autos	1	2	0	0	0	2	6	93	0	1	2	3	3	2	3	2	3	32	113	1
rec.motorcycles	1	2	0	1	0	1	1	19	87	0	0	3	0	2	0	2	3	26	133	3
rec.sport.baseball	2	3	0	0	0	1	0	0	0	152	12	2	0	2	0	3	3	15	51	2
rec.sport.hockey	0	0	0	0	0	0	0	0	0	0	194	1	0	0	0	1	0	5	29	1
sci.crypt	1	1	1	0	0	3	0	0	0	0	0	197	0	0	0	2	2	2	24	0
sci.electronics	0	18	0	5	1	6	0	1	1	0	1	53	79	1	7	4	3	12	50	2
sci.med	6	5	0	0	0	0	0	0	0	1	1	3	0	157	2	9	1	22	46	3
sci.space	5	3	0	0	0	2	1	0	0	0	1	2	1	1	156	6	1	10	56	1
soc.religion.christian	2	1	0	0	0	0	0	0	0	1	0	1	0	0	0	221	0	12	11	3
talk.politics.guns	1	0	0	0	0	0	0	0	0	0	1	2	0	0	0	4	105	14	120	2
talk.politics.mideast	2	0	0	0	0	1	0	0	1	0	0	1	0	0	0	2	2	260	12	0
talk.politics.misc	1	1	0	0	0	0	0	0	0	0	1	2	0	0	1	4	5	28	214	2
talk.religion.misc	51	1	0	0	0	0	0	0	0	0	1	1	0	0	0	61	10	8	55	48

Based on the Confusion Matrix, the newsgroups that most confuse the classifier are rec.motorcycle and talk.politics.misc, with rec.motorcycle being misclassified as talk.politics.misc 133 times.

This confusion was not expected since I would assume that motorcycle discussions and political discussions are distinct topics. However, given that the text data often contains subtle nuances, the model might incorrectly associate specific political discussions (especially those related to transport or motorist rights) with the "rec.motorcycles" class.

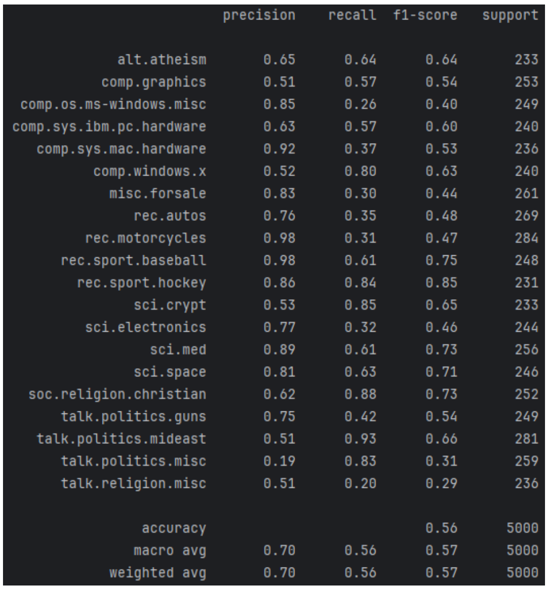
4. **<u>Based on the confusion matrix, report the individual accuracy scores for each class (1 mark).</u>**



```
alt.atheism: 0.0298
comp.graphics: 0.029
comp.os.ms-windows.misc: 0.0128
comp.sys.ibm.pc.hardware: 0.0274
comp.sys.mac.hardware: 0.0174
comp.windows.x: 0.0386
misc.forsale: 0.0156
rec.autos: 0.0186
rec.motorcycles: 0.0174
rec.sport.baseball: 0.0304
rec.sport.hockey: 0.0388
sci.crypt: 0.0394
sci.electronics: 0.0158
sci.med: 0.0314
sci.space: 0.0312
soc.religion.christian: 0.0442
talk.politics.guns: 0.021
talk.politics.mideast: 0.052
talk.politics.misc: 0.0428
talk.religion.misc: 0.0096
```

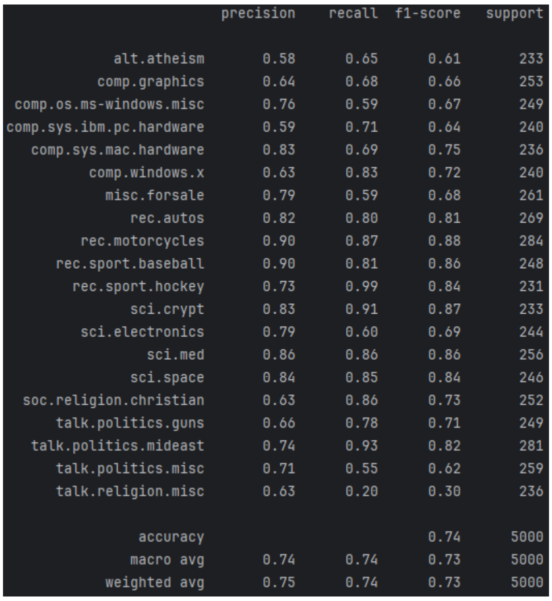
5. **Train a Complement Naive Bayes classifier and compare its classification results with Multinomial Naive Bayes (1 mark).**

Multinomial Naïve Bayes classifier classification results:



```
                          precision    recall  f1-score   support

             alt.atheism       0.65      0.64      0.64       233
           comp.graphics       0.51      0.57      0.54       253
 comp.os.ms-windows.misc       0.85      0.26      0.40       249
comp.sys.ibm.pc.hardware       0.63      0.57      0.60       240
   comp.sys.mac.hardware       0.92      0.37      0.53       236
          comp.windows.x       0.52      0.80      0.63       240
            misc.forsale       0.83      0.30      0.44       261
               rec.autos       0.76      0.35      0.48       269
         rec.motorcycles       0.98      0.31      0.47       284
      rec.sport.baseball       0.98      0.61      0.75       248
        rec.sport.hockey       0.86      0.84      0.85       231
               sci.crypt       0.53      0.85      0.65       233
         sci.electronics       0.77      0.32      0.46       244
                 sci.med       0.89      0.61      0.73       256
               sci.space       0.81      0.63      0.71       246
  soc.religion.christian       0.62      0.88      0.73       252
      talk.politics.guns       0.75      0.42      0.54       249
   talk.politics.mideast       0.51      0.93      0.66       281
      talk.politics.misc       0.19      0.83      0.31       259
      talk.religion.misc       0.51      0.20      0.29       236

                accuracy                           0.56      5000
               macro avg       0.70      0.56      0.57      5000
            weighted avg       0.70      0.56      0.57      5000
```

Complement Naïve Bayes Classifier classification results



```
                          precision    recall  f1-score   support

             alt.atheism       0.58      0.65      0.61       233
           comp.graphics       0.64      0.68      0.66       253
 comp.os.ms-windows.misc       0.76      0.59      0.67       249
comp.sys.ibm.pc.hardware       0.59      0.71      0.64       240
   comp.sys.mac.hardware       0.83      0.69      0.75       236
          comp.windows.x       0.63      0.83      0.72       240
            misc.forsale       0.79      0.59      0.68       261
               rec.autos       0.82      0.80      0.81       269
         rec.motorcycles       0.90      0.87      0.88       284
      rec.sport.baseball       0.90      0.81      0.86       248
        rec.sport.hockey       0.73      0.99      0.84       231
               sci.crypt       0.83      0.91      0.87       233
         sci.electronics       0.79      0.60      0.69       244
                 sci.med       0.86      0.86      0.86       256
               sci.space       0.84      0.85      0.84       246
  soc.religion.christian       0.63      0.86      0.73       252
      talk.politics.guns       0.66      0.78      0.71       249
   talk.politics.mideast       0.74      0.93      0.82       281
      talk.politics.misc       0.71      0.55      0.62       259
      talk.religion.misc       0.63      0.20      0.30       236

                accuracy                           0.74      5000
               macro avg       0.74      0.74      0.73      5000
            weighted avg       0.75      0.74      0.73      5000
```

Based on the classification results, the Complement Naïve Bayes Classifier is generally more accurate than the Multinomial Naïve.

Explanation:

- **Complement Naïve Bayes (CNB):** Typically performs better in scenarios with imbalanced datasets, which may explain its higher accuracy and better performance in this comparison.
- **Multinomial Naïve Bayes (MNB):** While also effective, MNB may not perform as well in some instances, especially when dealing with imbalanced classes.