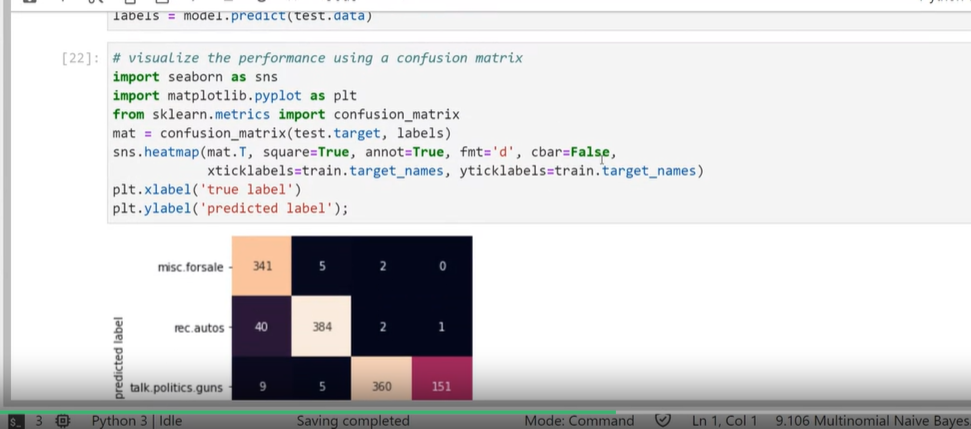
**Evaluation Methodology**

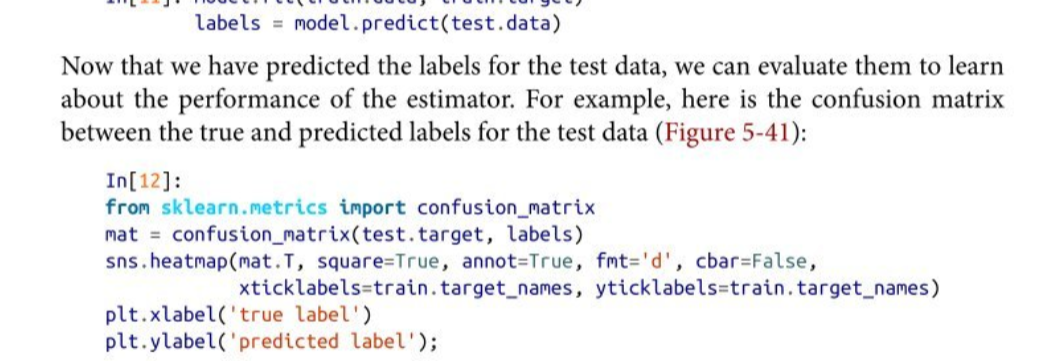
* Train model on following datasets:
  + Original split, simple tokenization and word counts, evaluate on validation set
  + Then do negation and compare
  + Then remove stop words and compare
* What if we predict the majority class the whole time (calculate predicting neutral class every time)? Accuracy can be close to the trivial classifier (accuracy = ratio of correct predictions to total predictions). For an unbalanced dataset, we can achieve high accuracy by simply predicting the majority label every time (doing nothing but returning the most popular class, neutral sentiment).
* Unbalanced data sets
* **Confusion matrix:** compares predicted values against ground truth or actual labels. It can be used to calculate a wide range of metrics, and to compare predicted values against actual values allowing more fine-grained evaluation.
* **Accuracy** = (TP + TN) / (TP + TN + FP + FN)
* **Recall** = TP / (TP + FN),  *or what propotion of x class did we actually identify and predict?*
* **Precision** = TP / (TP + FP)
* These metrics are more effective at illustrating the **performance on an umbalanced dataset**, and they **provide a baseline against which we can compare the performance of future iterations of our system**, where we might try different algorithms and different approaches. These metrics are a more principled approach to NLP.
* Confusion matrix allows for a more detailed breakdown of which classes the model is performing worse on, so for instance, to see if the performance breaks down when it comes to “mixed” sentiment classes.
* Do the same for cross-validation for new split
* Choose model and explain why 🡪 set model hyperparameters 🡪 configure the data 🡪 fit the model to the data 🡪 apply model to unseen data

**Perkins, J.** [**Python 3 text processing with NLTK 3 cookbook**](https://go.oreilly.com/university-of-london-worldwide/library/view/python-3-text/9781782167853/)**. (Birmingham: Packt Publishing Ltd, 2014). Chapter 7, pp.201-213.**

* “However, I find precision and recall to be much more useful metrics by themselves, as the F-measure can obscure the kinds of imbalances we saw with the NaiveBayesClassifier class.”

**Confusion Matrix Code**

* Plots the pairwise possibilities for each class (real and actual labels). Can see where the model made mistakes.



**Stratified K-Fold Cross-Validation**

From <https://www.linkedin.com/pulse/stratified-k-fold-cross-validation-in-depth-look-yeshwanth-n/>

### “What is Stratified K-Fold Cross-Validation?

Stratified K-Fold Cross-Validation is a variation of K-Fold Cross-Validation that ensures each fold maintains the same proportion of observations for each target class as the complete dataset. This is especially crucial for datasets where one class might be heavily underrepresented.

**Why Use Stratified K-Fold Cross-Validation?**

1. **Balanced Splits:** Standard K-Fold Cross-Validation can sometimes result in imbalanced class distributions across the folds. Stratified K-Fold ensures every fold has a representative class distribution.
2. **Better Generalization:** Since each fold has a similar distribution, the model is less likely to be biased towards any particular class, leading to better generalization on unseen data.
3. **Robust Performance Metrics:** With balanced class distribution in each fold, performance metrics such as accuracy, precision, and recall become more reliable.

“

* When to use macro and micro-average:
  + <https://www.evidentlyai.com/classification-metrics/multi-class-metrics#:~:text=Macro%2Daveraging%20gives%20equal%20weight,same%20and%20identical%20to%20accuracy>.
  + <https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f>
  + <https://safjan.com/micro-and-macro-averages-in-multiclass-multilabel-problems/>

“**Micro averages** are useful when the **classes are imbalanced** and it is important to have a better understanding of the model's **performance on the majority class**.

**Macro averages,** on the other hand, are useful when **all classes are of equal importance** and you want to have a better understanding of the model's **performance on each class individually**.

In some cases, it may be useful to report both micro and macro averages to get a more comprehensive understanding of the model's performance.”

<https://www.geeksforgeeks.org/stratified-k-fold-cross-validation/>

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html>

<https://medium.com/@juanc.olamendy/a-comprehensive-guide-to-stratified-k-fold-cross-validation-for-unbalanced-data-014691060f17>