Authorship Attribution

Shiqi Qu, Xinxuan Wei, Zhihui Hong
IST 736 Text Mining M001, Prof Bei Yu

Research Question

- Which vectorizer is better for authorship attribution?
- Which classifier is more suitable for this task?

Corpus

- 5000 texts from 50 authors (100 per author) from the Reuters Corpus Volume 1
- 2500 texts for training; 2500 texts for testing
- Genre: newswire stories

Preprocessing

- Convert every digit of number into '@'

Vectorization

- Unigram term frequency
- Unigram TFidf
- Unigram + Bigram term frequency
- Unigram + Bigram TFidf
- Character 3gram
- Parameter: stop words, minimum term frequency, lowercase, max_feature

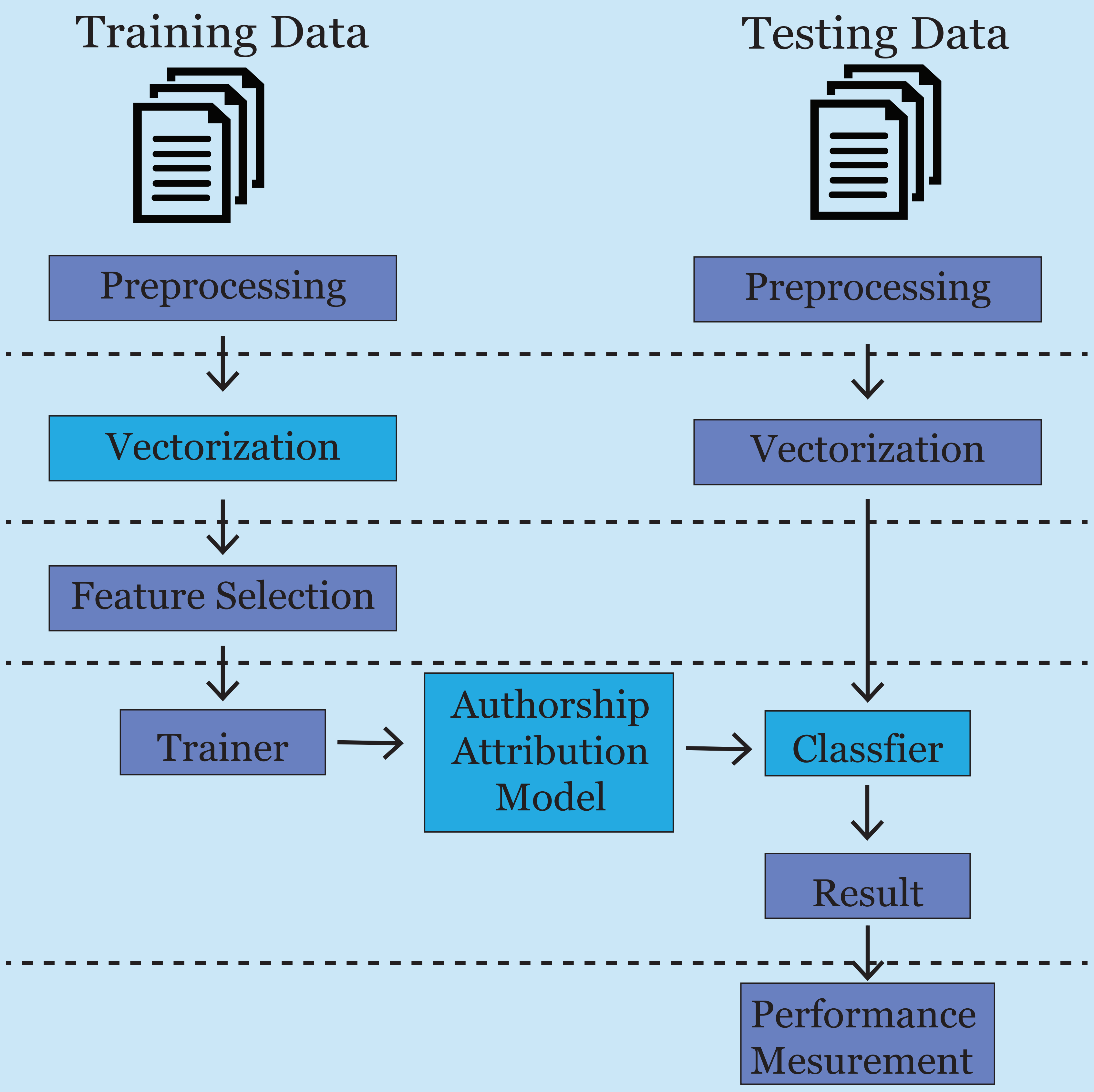
Classifier

- Multinomial Naïve Bayes
- Support Vector Machine
- K Nearest Neighbor

Evaluation

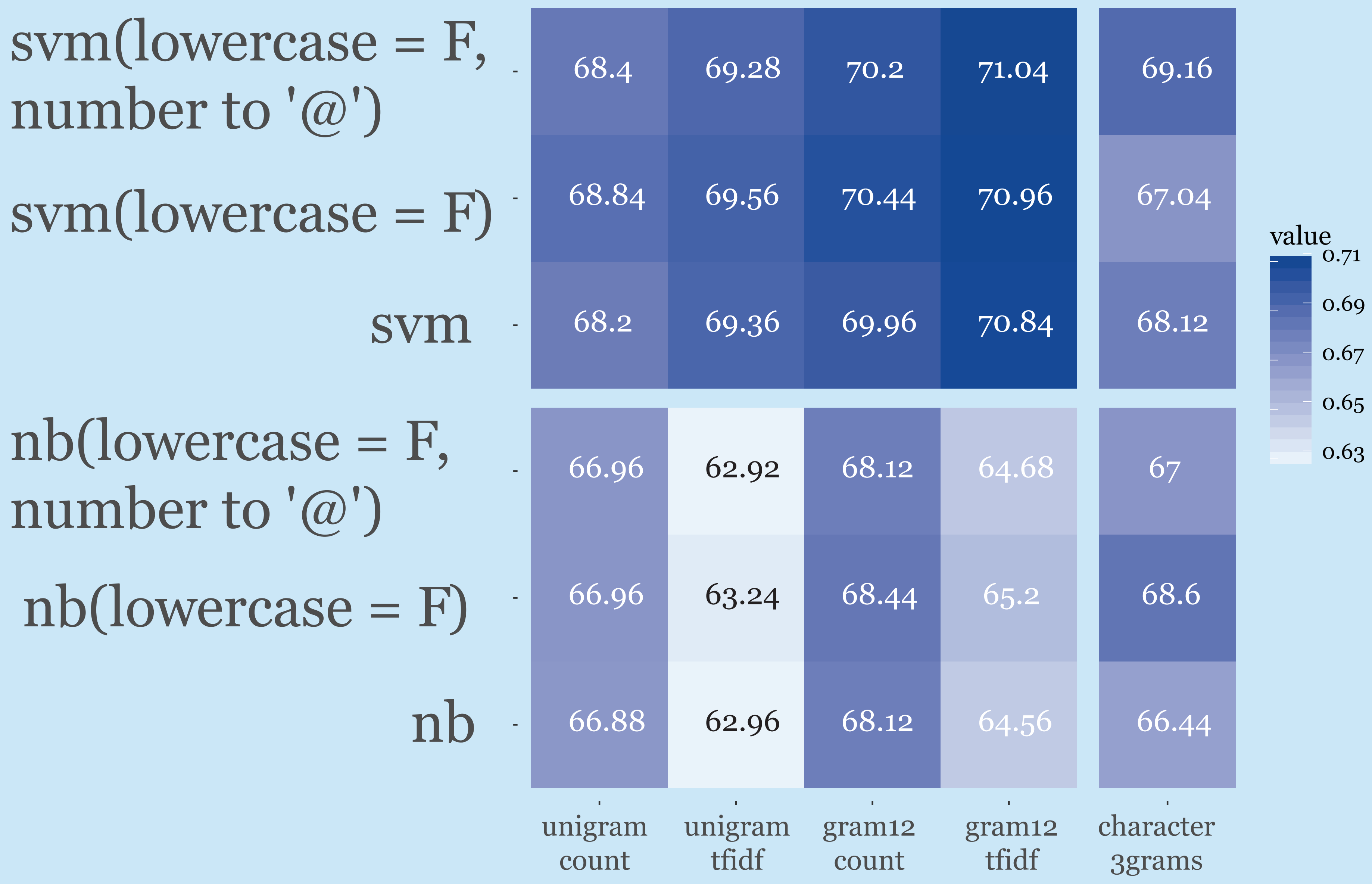
- Accuracy, precision, recall, confusion matrix

Workflow



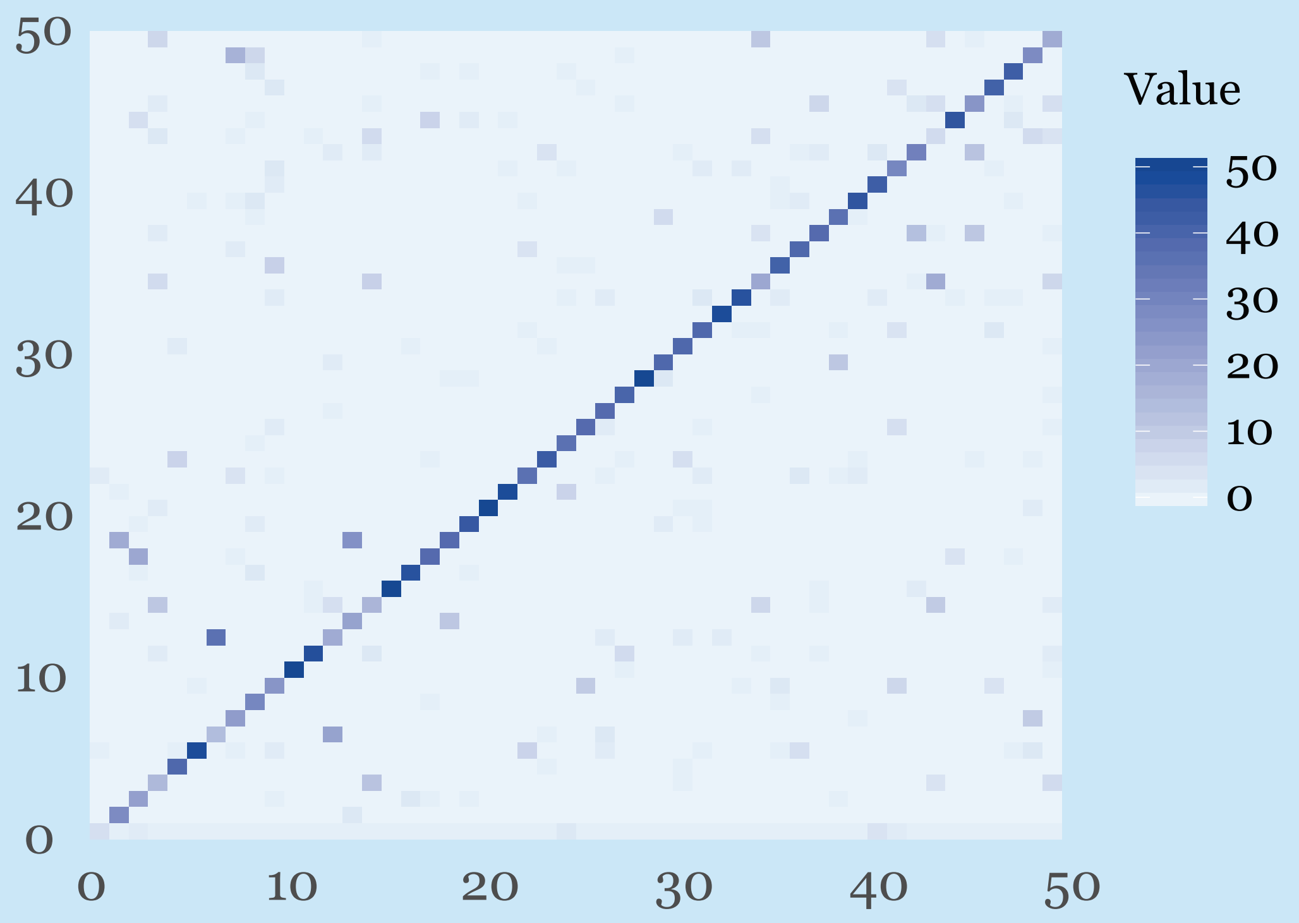
Results

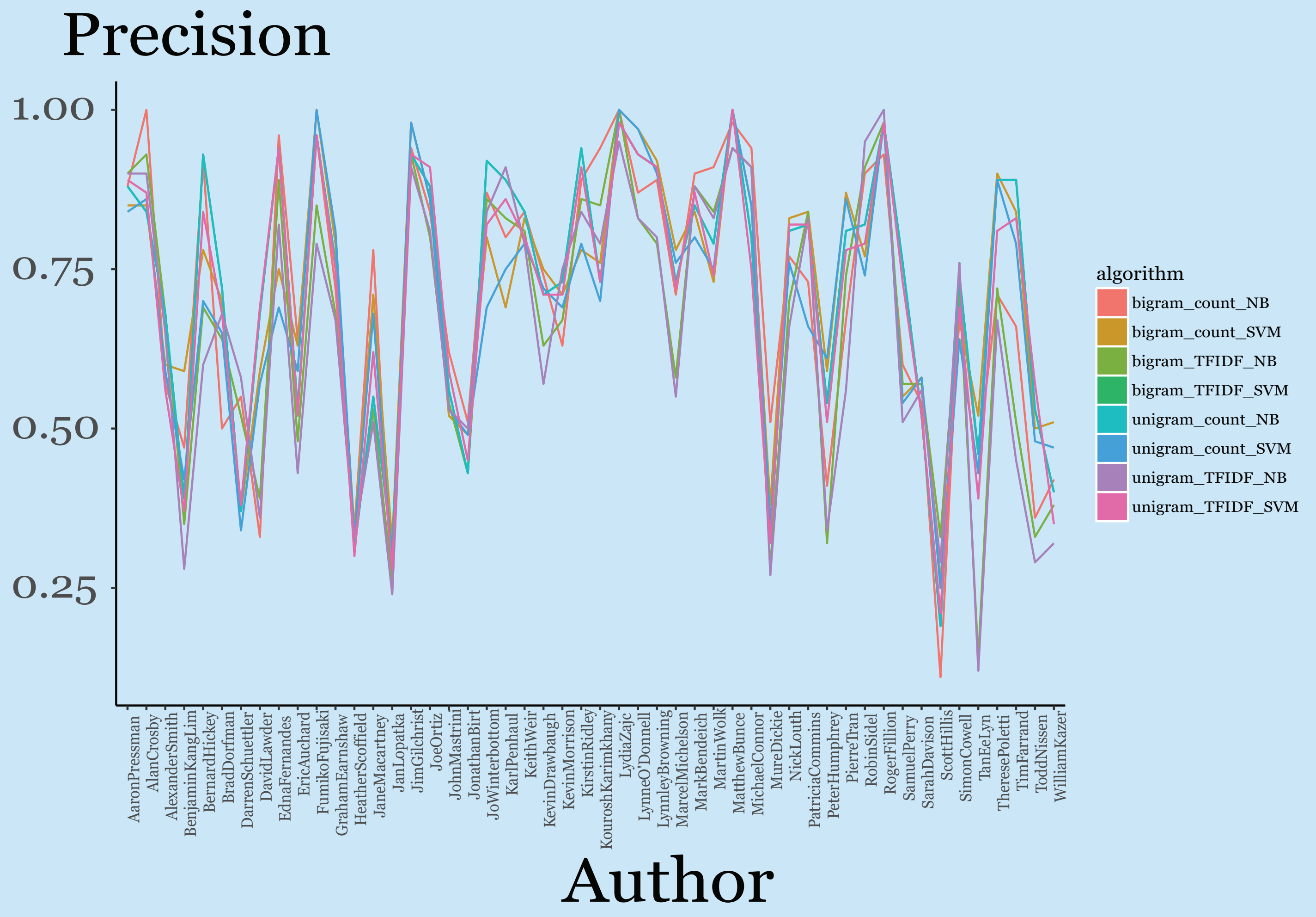
(Units: %)

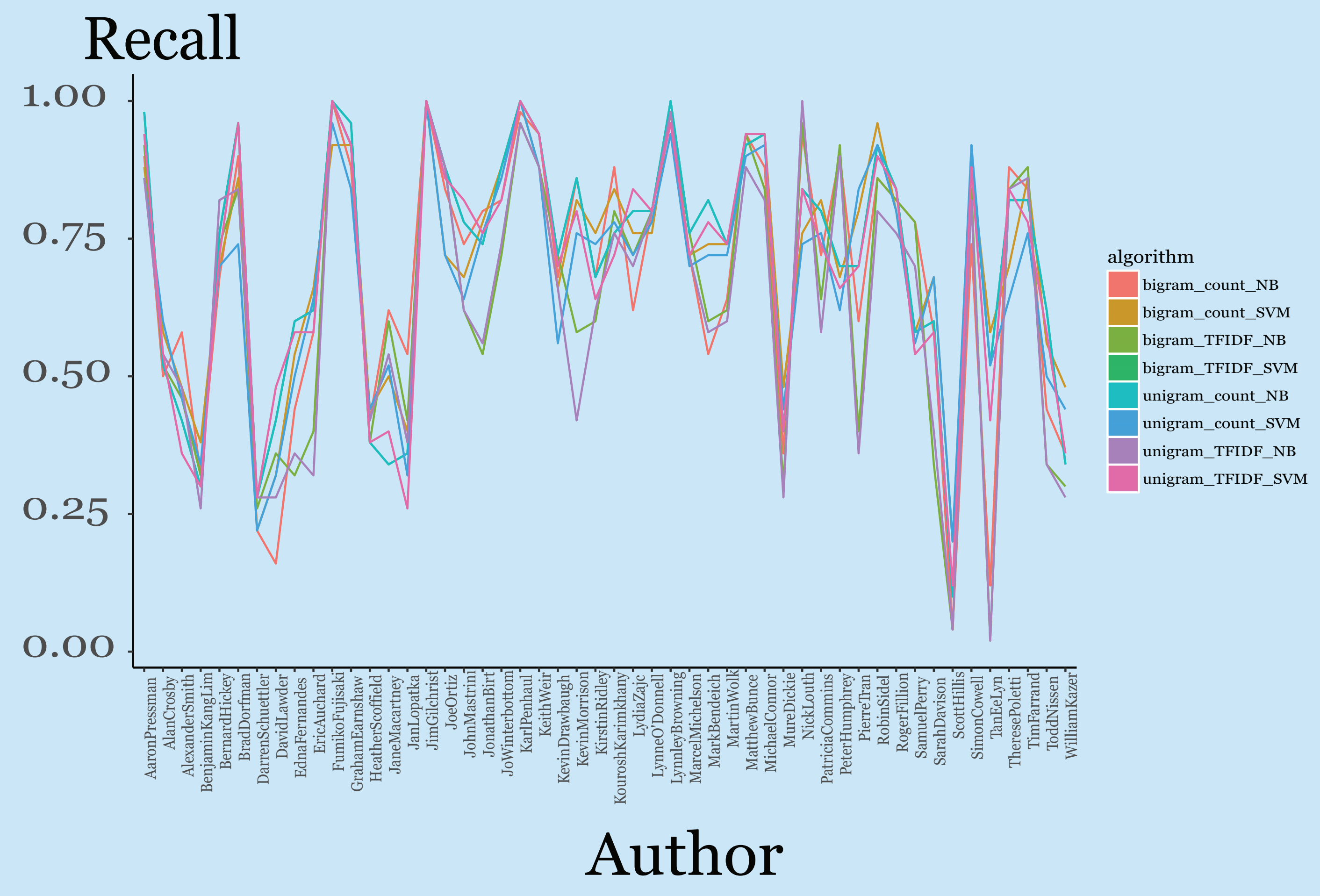


	unigram count	unigram tfidf	gram12 count	gram12 tfidf	character 3grams
svm(lowercase = F, number to '@')	68.4	69.28	70.2	71.04	69.16
svm(lowercase = F)	68.84	69.56	70.44	70.96	67.04
svm	68.2	69.36	69.96	70.84	68.12
nb(lowercase = F, number to '@')	66.96	62.92	68.12	64.68	67
nb(lowercase = F)	66.96	63.24	68.44	65.2	68.6
nb	66.88	62.96	68.12	64.56	66.44

Confusion Matrix







Conclusion

- Unigram and bigram with TFidf normalization can best represent texts for authorship attribution.
- Performance of classifiers: SVM > MNB > KNN
- SVM suits for the task since it consistently achieves good performance for the identification tasks.
- There is no need to select specific features. Preprocessing and weighting of features is not critical since it leads to identical results.
- Accuracy decreased if converting words to lowercase. Possible reason is some authors like to use short sentences which may generate more words with capital letters.