Text Categorization Report

1. Introduction:

This report presents the results of a text categorization task performed on the 20 Newsgroups dataset. The goal was to classify news articles into 20 different categories using various machine learning classifiers and feature extraction methods.

2. Methods:

We evaluated three classifiers on this multi-class classification task:

- Naive Bayes
- Logistic Regression
- Support Vector Machine (SVM)

For feature extraction, we compared three methods:

- Count Vectorizer (word counts)
- Term Frequency Vectorizer (binary occurrence of words)
- TF-IDF Vectorizer (term frequency-inverse document frequency)

Additionally, for the best performing classifier, we performed hyperparameter tuning on the following parameters:

- Lowercasing
- Stop Words
- ngram_range
- max_features

3. Results:

3.1 Classifier and Feature Comparison:

From the initial comparison, SVM with TF-IDF Vectorizer achieved the best performance:

- Accuracy: 0.72
- Macro F1-Score: 0.71
- Weighted F1-Score: 0.72

Based on this, we selected SVM with TF-IDF for further experiments.

3.2 Results Table:
The following table summarizes the Precision, Recall, and F1 scores for each classifier and feature extraction method:

Classifier	Feature Extraction	Precision	Recall	F1-Score	Accuracy
Naive Bayes	Count Vectorizer	0.69	0.60	0.58	0.60
	Term Frequency Vectorizer	0.72	0.60	0.60	0.62
	TF-IDF Vectorizer	0.78	0.65	0.65	0.65
Logistic Regression	Count Vectorizer	0.66	0.65	0.66	0.65
	Term Frequency Vectorizer	0.66	0.65	0.65	0.66
	TF-IDF Vectorizer	0.73	0.72	0.71	0.72
SVM	Count Vectorizer	0.52	0.49	0.50	0.49
	Term Frequency Vectorizer	0.54	0.51	0.51	0.51
	TF-IDF Vectorizer	0.73	0.72	0.72	0.72
SVM with TF-IDF	Lowercase=True	0.73	0.72	0.72	0.72
	Lowercase=False	0.72	0.70	0.71	0.71
	Stop Words (english)	0.75	0.73	0.74	0.74
	ngram_range (1, 2)	0.74	0.71	0.71	0.72
	max_features=5000	0.67	0.66	0.66	0.66

3.3 Best Model:
The best-performing configuration was SVM with TF-IDF Vectorizer and stop_words='english', which yielded the following results:

- Accuracy: 0.74
- Macro F1-Score: 0.73
- Weighted F1-Score: 0.74

4. Discussion:
The SVM classifier with TF-IDF and stop word removal performed the best. Removing common stop words allowed the model to focus on more informative terms, improving accuracy and F1-scores. Lowercasing and adding bigrams had little effect while limiting the vocabulary with max_features decreased performance.

Future work could explore additional techniques like feature selection, and further tuning of SVM's hyperparameters, such as the regularization parameter (C) and kernel functions.

5. Conclusion:

- Best Model: SVM with TF-IDF Vectorizer and stop_words='english'.

- Final Performance:
    - Accuracy: 0.74
    - Macro F1-Score: 0.73
    - Weighted F1-Score: 0.74

The results indicate that SVM is an effective classifier for text categorization tasks, especially when combined with TF-IDF feature extraction and stop word removal.