Reproducing Pang & Lee (2002) on the Movie-Reviews Corpus

Debei — July 2025

1 Corpus & Experimental Setup

Item	**Specification**
Corpus	`movie_reviews` (1 000 pos / 1 000 neg)
Fold scheme	`cvNNN_%10` $\Rightarrow$ 10 folds, each 100 pos / 100 neg
Text prep	lower-case; collapse whitespace; *no* stemming; punctuation as token delimiters
Hardware	Colab CPU or local machine
Software	Python 3.11, scikit-learn , NLTK , pandas , numpy

2 Baseline Results (Original Settings)

Model	Vectoriser	Hyper-parameters	Accuracy
Naïve Bayes	Unigram presence	$\alpha = 1$	0.821
MaxEnt (LogReg)	Unigram presence	$C = 10\,000$	0.865
Linear SVM	Unigram presence	$C = 1$	0.849