Three text classification studies

French doctoral theses

We grouped 259,774 thesis records into 12 disciplines and an “Other” category. The classifier used TF-IDF, truncated singular value decomposition, linear discriminant analysis and naïve Bayes on titles and descriptions.

Held-out test set	Accuracy	Macro-recall	Macro-F1
Majority-class baseline	73.3%	7.7%	6.5%
LDA + naïve Bayes	45.3%	50.4%	36.9%

Both models used the same 51,949 records from a stratified 80/20 split. The trained model achieved a higher macro-F1. The majority baseline had higher accuracy because 73.3% of the test set belonged to “Other”.

Emotions in English comments

We compared tokenization, stemming and lemmatization on 5,937 comments labeled anger, fear or joy. We fit the TF-IDF vocabulary on each training fold and evaluated naïve Bayes over five stratified folds.

Mean macro-F1 was 0.319 with tokenization, 0.351 with stemming and 0.332 with lemmatization. Stemming also produced the highest mean accuracy at 37.5%.

Entity sentiment in tweets

The dataset supplied 74,682 training rows and 1,000 validation rows across four labels. Linear discriminant analysis on TF-IDF features gave 49.8% validation accuracy and a Cohen’s kappa of 0.317.

The validation vocabulary came from training, but inverse document frequencies were recalculated on the validation corpus. The score describes that specific evaluation.

Sources: team model metrics, cross-validation output and rendered sentiment report.