Comparative Analysis of Machine Learning Models for Text Classification

Abstract

This report compares RandomForest, XGBoost, and Logistic Regression for text classification using Bag-of-Words (BoW) and TF-IDF features. Performance was evaluated with accuracy, F1-score, and confusion matrices. Logistic Regression achieved the best results, showing that simple linear models can outperform complex ensembles in this task.

1 Introduction

Text classification is a core NLP task used in sentiment analysis, spam detection, and topic labeling. This study evaluates three classifiers with two feature extraction techniques to determine the most effective approach.

2 Methodology

A dataset of text and titles was preprocessed, cleaned, and sampled to 100k entries for training and testing. Features were extracted using:

- **BoW**: word frequency representation
- **TF-IDF**: weighted frequency representation

The classifiers compared were RandomForest, XGBoost, and Logistic Regression. Performance was measured using accuracy, F1-score, and OOB score (for RandomForest).

3 Results

Table 1 summarizes the performance of all models. Logistic Regression outperformed others with ∼87.8% accuracy and F1-score.

Table 1: Summary of Model Performance

Model	**Features**	**Accuracy**	**F1-score**	**OOB Score**
RandomForest	BoW	0.8608	0.8616	0.8459
RandomForest	TF-IDF	0.8621	0.8634	0.8469
XGBoost	BoW	0.8513	0.8526	N/A
XGBoost	TF-IDF	0.8499	0.8507	N/A
Logistic Regression	BoW	0.8776	0.8769	N/A
Logistic Regression	TF-IDF	0.8776	0.8769	N/A

4 Challenges

During the study, several challenges were observed:

1. **High dimensionality:** Text features created extremely large sparse matrices, increasing computation time.
2. **Label inconsistencies:** Mapping of target labels caused confusion, especially in XGBoost results.
3. **Sampling trade-off:** Reducing dataset size ensured feasibility but may have slightly limited model performance.
4. **Feature similarity:** BoW and TF-IDF gave nearly identical results, showing limited benefit from more complex weighting.

5 Conclusion

Logistic Regression with BoW or TF-IDF delivered the best results, outperforming RandomForest and XGBoost. For this dataset, simple linear models proved more effective than complex ensembles, showing that frequency-based features are strong predictors for text classification.