Sports vs Politics Classification Report

Aditya Arun Kumar Yadav (M25CSA001)

Github Link - https://github.com/yadityax/NLU-Assignment

Abstract

This report presents binary text classification system for distinguishing sports from politics news. Five ML models were compared using BoW, TF-IDF, and n-grams on 2,000 articles (1,000 per category) with a proper train/validation/test split (70:10:20). Linear SVC with TF-IDF bigrams achieved best test performance at 95.5% accuracy and 92.5% validation accuracy.

1. Introduction and Objectives

In this report we perform binary classification of news articles as **Sports** or **Politics** using Kaggle's News Category Dataset—a large-scale collection of HuffPost headlines spanning 2012-2022. The objectives are to:

(1) obtain and prepare data from Kaggle

(2) implement multiple feature representations (BoW, TF-IDF, n-grams)

(3) compare five ML models

(4) evaluate performance

(5) analyze limitations

2. Data Collection and Description

Dataset Source: News Category Dataset from Kaggle by Rishabh Misra (https://www.kaggle.com/datasets/rmisra/news-category-dataset)

Data Characteristics: Short texts combining headlines and descriptions. Sports articles cover football, basketball, cricket, Olympics with terms like *championship*, *victory*, *playoff*. Politics articles discuss elections, policy, legislation with terms like *president*, *congress*, *vote*,

election. The US-centric HuffPost content and short text format present realistic classification challenges.

Preprocessing: Dataset manually downloaded from Kaggle, filtered for sports/politics categories, balanced to 1,000 samples each, and saved as text files. Data split using 70:10:20 ratio for train/validation/test sets (1,400/200/400 samples) with stratification to maintain class balance.

3. Methodology

Feature Representations:

- Bag of Words (BoW) with bigrams (1-2): Word frequency vectors including unigrams and bigrams, max 5000 features
- TF-IDF with bigrams (1-2): Term Frequency-Inverse Document Frequency weighting, reduces common word impact, max 5000 features
- TF-IDF with trigrams (1-3): Extends to three-word phrases to test longer n-gram benefit, max 5000 features

Machine Learning Models:

- Multinomial Naive Bayes (BoW 1-2) - Probabilistic classifier assuming feature independence, fast training
- Logistic Regression (BoW 1-2) - Linear discriminative model with regularization
- Linear SVC (TF-IDF 1-2)- Support Vector Classifier maximizing margin, robust to high dimensions
- Random Forest (TF-IDF 1-2) - Ensemble of 100 decision trees with parallel processing
- Logistic Regression (TF-IDF 1-3)- Tests trigram impact on performance

Evaluation: Three-way split strategy (70:10:20) enables proper model selection. Accuracy, precision, recall, and F1-score computed on both validation (200 samples) and test (400 samples) sets with "sports" as positive class. Validation set used for model comparison and selection, test set reserved for final unbiased performance assessment.

4. Results and Analysis

Model Performance:

Validation Set (200 samples):

Model	Accuracy	Precision	Recall	F1
MultinomialNB (BoW 1-2)	0.9246	0.9293	0.9200	0.9246
LogisticRegression (BoW 1-2)	0.9095	0.8942	0.9300	0.9118
LinearSVC (TF-IDF 1-2)	0.9246	0.9293	0.9200	0.9246
RandomForest (TF-IDF 1-2)	0.8643	0.8544	0.8800	0.8670
LogisticRegression (TF-IDF 1-3)	0.9246	0.9208	0.9300	0.9254

Test Set (400 samples):

Model	Accuracy	Precision	Recall	F1
MultinomialNB (BoW 1-2)	0.9476	0.9497	0.9450	0.9474
LogisticRegression (BoW 1-2)	0.9327	0.9179	0.9500	0.9337
LinearSVC (TF-IDF 1-2)	0.9551	0.9550	0.9550	0.9550
RandomForest (TF-IDF 1-2)	0.9027	0.8966	0.9100	0.9032
LogisticRegression (TF-IDF 1-3)	0.9352	0.9394	0.9300	0.9347

Key Findings: Linear SVC with TF-IDF bigrams achieved best test accuracy (95.5%) with strong validation performance (92.5%), demonstrating robust discrimination despite short text (10-30 words). Multinomial Naive Bayes closely followed (94.8% test) with fastest training. All models exceeded 90% test accuracy.

Analysis:

- TF-IDF vs BoW: TF-IDF outperformed raw counts (95.5% vs 93.3% test accuracy for Logistic Regression), confirming better feature weighting
- Validation consistency: LinearSVC showed consistent val→test performance (92.5%→95.5%), indicating good generalization without overfitting
- Bigrams sufficiency: Trigrams showed no improvement over bigrams on short text, suggesting diminishing returns
- Linear superiority: Linear models (SVC, Logistic Regression, Naive Bayes) outperformed Random Forest (90.3% test) on sparse high-dimensional text
- Three-way split benefit: Validation set (10%) enabled unbiased model selection before final test evaluation, following ML best practices
- Short text success: 95.5% test accuracy on headlines demonstrates distinct vocabularies between sports and politics despite limited context

The ~4.5% test error rate (18 misclassifications) indicates minimal overlap between categories. Validation-test consistency confirms robust model performance.

5. Limitations

Dataset: (1) Short text only (headlines+descriptions) lacks full article context, (2) Single source bias (HuffPost) limits generalization to other outlets, (3) US-centric content may not transfer internationally, (4) 2012-2022 temporal range may not reflect current language trends, (5) Imbalanced original data required downsampling politics (35K→1K).

Features: (1) BoW/TF-IDF ignore word order and semantics, (2) Fixed vocabulary cannot handle new terms/entities, (3) Sparse 5000-dimensional vectors, (4) No contextual understanding beyond n-gram statistics.

Models: (1) Entity dependence on specific team/politician names rather than topics, (2) Generalization uncertainty across different news sources, (3) Short text sensitivity to noise and typos, (4) Potential category overlap (sports politics coverage).

6. Conclusion

This study successfully classified HuffPost news articles as sports or politics using Kaggle's News Category Dataset (209K articles, 42 categories), achieving 95.5% test accuracy with Linear SVC on 2,000 balanced samples using rigorous 70:10:20 train/validation/test split. Five models compared across BoW and TF-IDF representations demonstrated that:

(1) Linear SVC with TF-IDF bigrams performs best (95.5% test, 92.5% validation)

(2) TF-IDF outperforms raw BoW

(3) Bigrams are sufficient—trigrams add no value on short text

(4) Linear classifiers beat ensemble methods on sparse text

(5) Validation set enables proper model selection before test evaluation

(6) Strong performance (>95%) possible despite short text (10-30 words).