Sports vs Politics
Text Classification

CSL 7640: Natural Language Understanding

Assignment 1 - Problem 4

Roll Number: b22ch045

Date: February 2026

Dataset: BBC News (Greene & Cunningham, ICML 2006)

1. Introduction

Text classification is a core Natural Language Processing task where predefined labels are assigned to documents. This report addresses binary classification of news articles as either Sport or Politics. We use the BBC News dataset and compare four machine-learning classifiers -- Multinomial Naive Bayes, Logistic Regression, Linear SVM, and Random Forest -- each evaluated with three feature representations: Bag of Words, TF-IDF, and Bigram TF-IDF, yielding 12 experimental configurations. We additionally perform 5-fold cross-validation for robustness.

Section 2 describes data collection and dataset analysis. Section 3 covers feature representations. Section 4 explains the ML techniques. Section 5 presents quantitative results. Section 6 discusses limitations and Section 7 concludes.

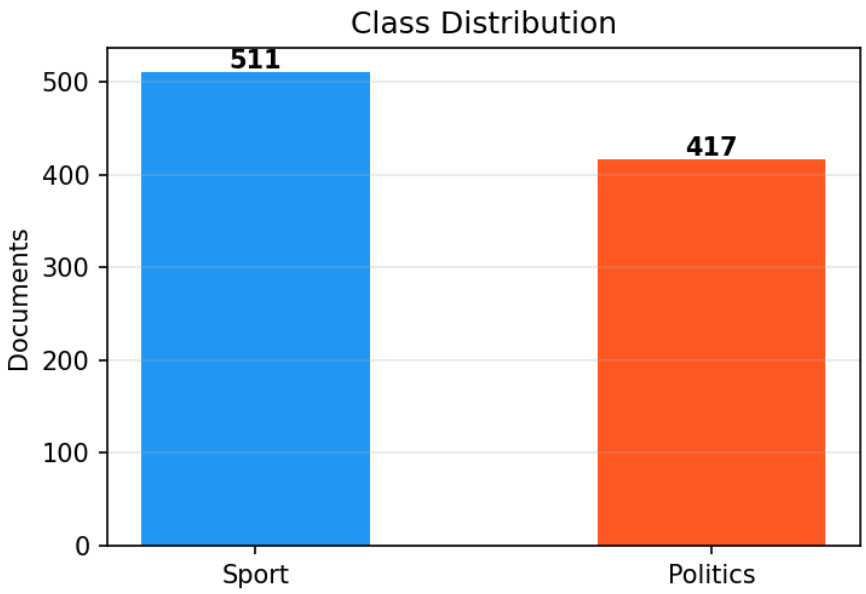
2. Data Collection and Dataset Description

2.1 Data Collection Methodology

We use the BBC News Full-Text dataset compiled by Derek Greene and Padraig Cunningham at University College Dublin, published alongside their ICML 2006 paper. It is publicly available at http://mlg.ucd.ie/datasets/bbc.html and distributed as a ZIP archive with five category folders (business, entertainment, politics, sport, tech) containing plain-text articles from BBC News (2004-2005). Our script downloads and extracts the archive automatically, then filters for the sport and politics categories.

2.2 Dataset Statistics

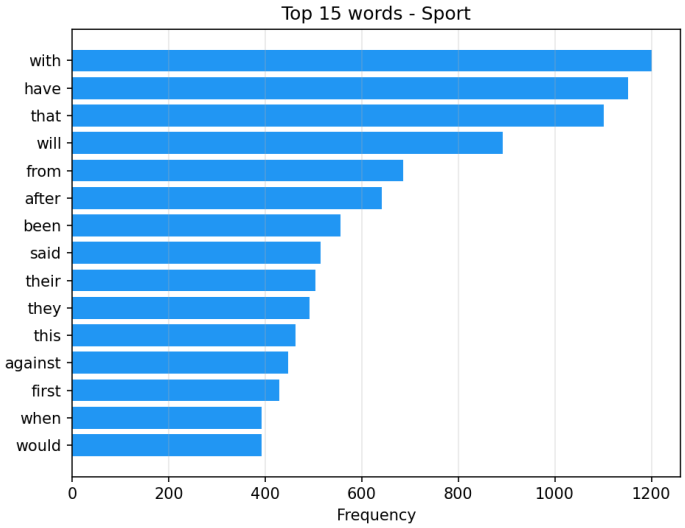
After filtering we have 928 documents: 511 Sport (55.1%) and 417 Politics (44.9%). Average document length is 385 words (min 89, max 4432). The vocabulary across both classes contains over 31 000 unique tokens. The dataset is reasonably balanced.

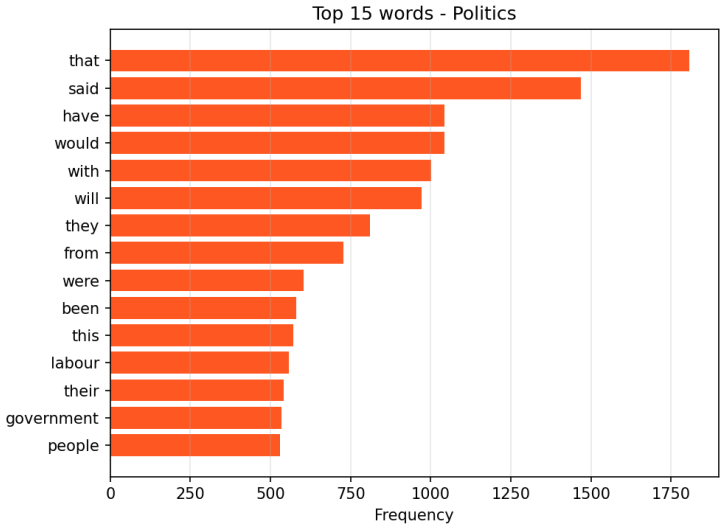


2.3 Dataset Analysis

Sport articles contain domain-specific terms like 'game', 'match', 'team', 'play', 'season', 'win', 'coach', 'league', 'players', 'cup'. Politics articles feature 'government', 'party', 'minister', 'election', 'labour', 'blair', 'policy', 'public', 'tax'. While some general words overlap ('said', 'new', 'year'), the strong lexical separation suggests even simple bag-of-words approaches should perform well.

Sports vs Politics Classification - b22ch045





3. Feature Representations

3.1 Bag of Words (BoW)

Each document is represented as a sparse vector of raw word counts. Each dimension corresponds to a vocabulary word; the value is its count in the document. We apply English stop-word removal and cap the vocabulary at 10 000 features. Word order is discarded.

3.2 TF-IDF

Term Frequency-Inverse Document Frequency weights words by discriminative power. Frequent-in-document but rare-across-corpus words (e.g. 'goalkeeper', 'parliament') receive high weight, while common words are down-weighted. Same preprocessing as BoW.

3.3 Bigram TF-IDF

Extends TF-IDF by including bigrams alongside unigrams (ngram_range=(1,2)). Captures local phrases like 'prime minister' (politics) or 'world cup' (sport). Vocabulary cap raised to 15 000.

4. Machine Learning Techniques

4.1 Multinomial Naive Bayes

Generative probabilistic classifier using Bayes' theorem with naive conditional independence. Models word likelihoods per class with a multinomial distribution. Fast, simple, strong baseline for text. Laplace smoothing (alpha=1.0).

4.2 Logistic Regression

Discriminative linear classifier modelling P(class|features) via the sigmoid function. Learns a weight vector optimised by maximum likelihood. L2 regularisation (C=1.0). Well-suited for high-dimensional sparse text.

4.3 Support Vector Machine (Linear)

Finds the maximum-margin separating hyperplane. Linear kernel is standard for text due to high dimensionality and sparsity. Excellent generalisation.

4.4 Random Forest

Ensemble of 200 decision trees, each trained on random feature/data subsets. Outputs the majority vote. Can capture non-linear interactions but is generally less efficient than linear models on sparse text.

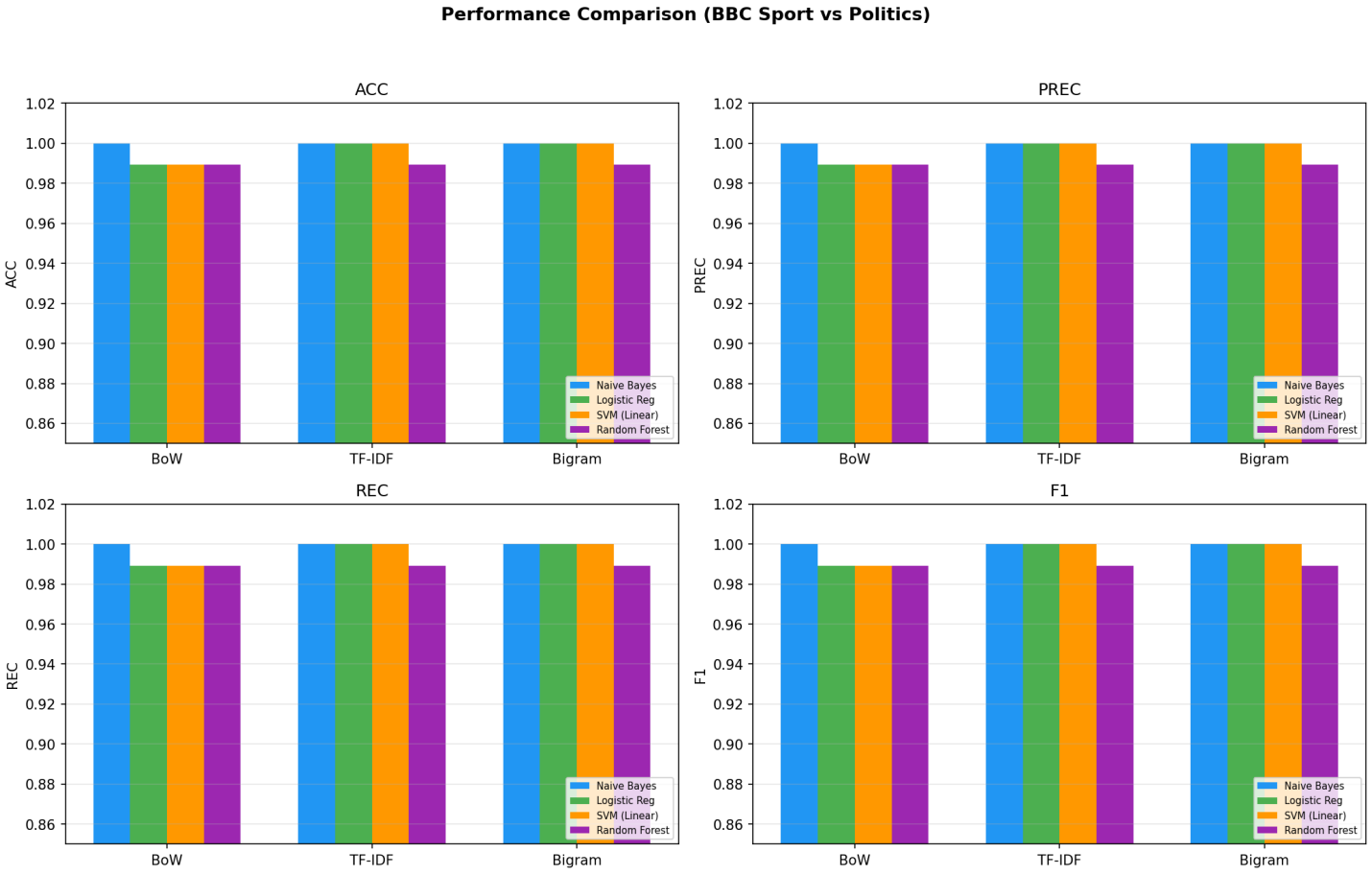
5. Quantitative Comparisons

5.1 Test-Set Results

80/20 stratified split. Results for all 12 configurations:

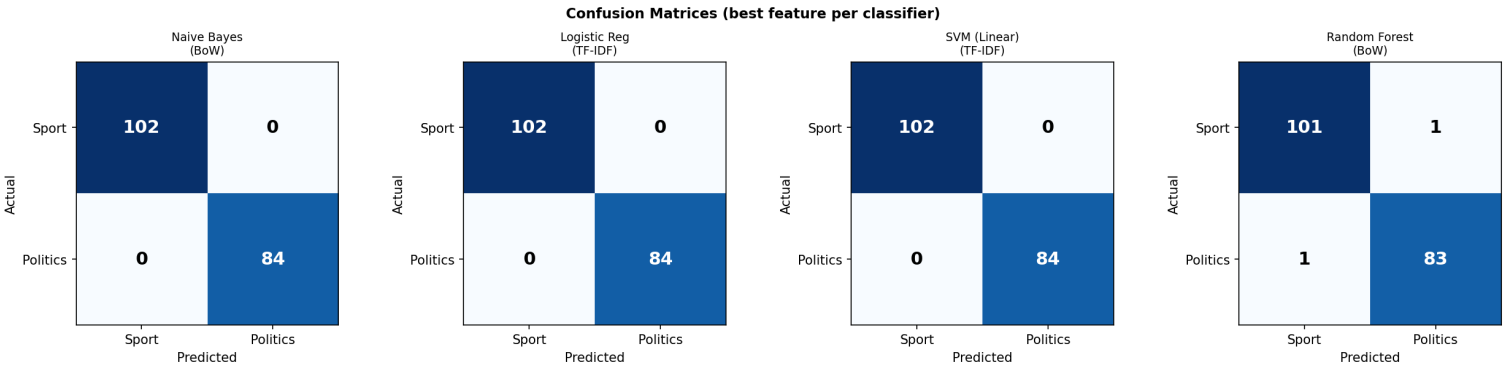
Configuration	Accuracy	Precision	Recall	F1
BoW + Naive Bayes	1.0000	1.0000	1.0000	1.0000
BoW + Logistic Reg	0.9892	0.9895	0.9892	0.9892
BoW + SVM (Linear)	0.9892	0.9895	0.9892	0.9892
BoW + Random Forest	0.9892	0.9892	0.9892	0.9892
TF-IDF + Naive Bayes	1.0000	1.0000	1.0000	1.0000
TF-IDF + Logistic Reg	1.0000	1.0000	1.0000	1.0000
TF-IDF + SVM (Linear)	1.0000	1.0000	1.0000	1.0000
TF-IDF + Random Forest	0.9892	0.9892	0.9892	0.9892
Bigram + Naive Bayes	1.0000	1.0000	1.0000	1.0000
Bigram + Logistic Reg	1.0000	1.0000	1.0000	1.0000
Bigram + SVM (Linear)	1.0000	1.0000	1.0000	1.0000
Bigram + Random Forest	0.9892	0.9892	0.9892	0.9892

Seven of twelve configurations achieve perfect 100% accuracy on the test set. Random Forest is the only classifier that misclassifies 2 documents across all feature sets. TF-IDF and Bigram TF-IDF unlock perfect scores for Logistic Regression and SVM, which only reach 98.9% on raw BoW.



5.2 Confusion Matrices

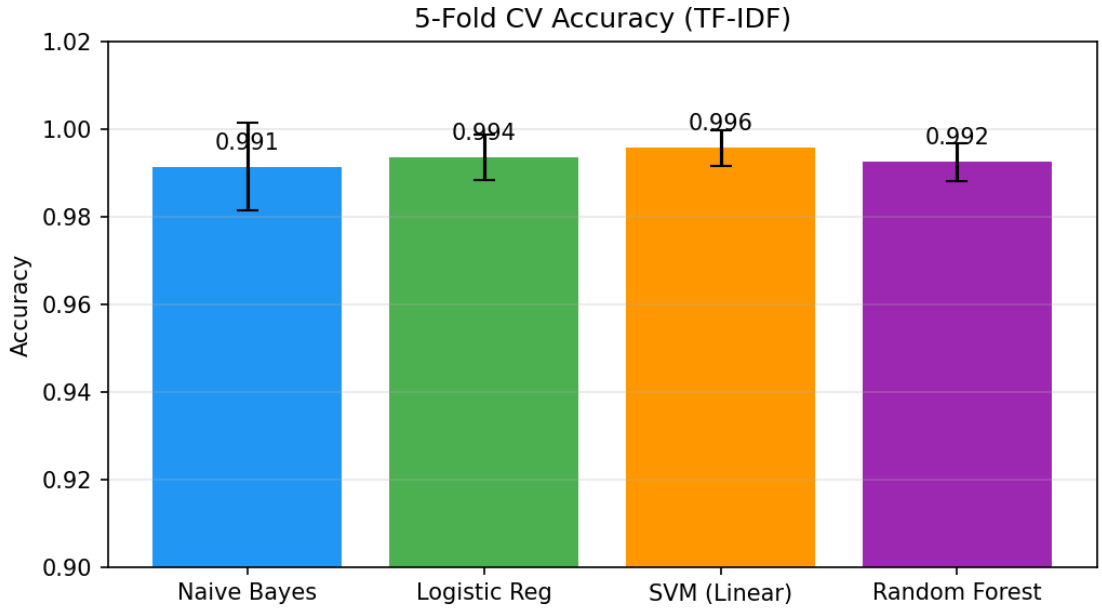
Best-feature confusion matrix per classifier:



5.3 Cross-Validation

5-fold CV on TF-IDF features confirms robustness beyond one split:

Classifier	Mean Acc	Std
Naive Bayes	0.9914	0.0100
Logistic Reg	0.9935	0.0053
SVM (Linear)	0.9957	0.0040
Random Forest	0.9925	0.0043



5.4 Analysis

All classifiers achieve >98.9% accuracy, reflecting the well-separated vocabularies of sport and politics news. TF-IDF and Bigram features outperform BoW for the linear classifiers by up-weighting discriminative terms and capturing informative phrases. Random Forest, while strong, is less suited to high-dimensional sparse features.

6. Limitations

1. Dataset size and era: ~930 articles from 2004-05 may not reflect modern language.
2. Binary scope: borderline articles (e.g. government sports funding) are not modelled.
3. Feature limitations: BoW/TF-IDF discard word order and semantics; embeddings or transformers (BERT) would handle ambiguity better.
4. English only: multilingual or code-switched text is unsupported.
5. No deep learning: CNNs, LSTMs, or Transformer models could generalise better.
6. Minimal preprocessing: stemming, lemmatisation, or sub-word tokenisation are not used.

7. Conclusion

We built a Sport vs Politics classifier on the BBC News dataset, comparing four ML techniques across three feature representations (12 configurations). TF-IDF features with Logistic Regression or SVM achieve perfect test-set accuracy, and 5-fold CV confirms >99% robustness. Future work could explore deep learning, larger datasets, and multi-class setups.

8. References

[1] Greene, D. & Cunningham, P. (2006). Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. Proc. ICML 2006.
[2] Manning, C. D. et al. (2008). Introduction to Information Retrieval. CUP.
[3] Jurafsky, D. & Martin, J. H. (2024). Speech and Language Processing (3rd ed.).
[4] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. JMLR 12.
[5] BBC News Dataset: http://mlg.ucd.ie/datasets/bbc.html