Project Report: Text Classification of News Articles (Sports vs. Politics)

Pragati Rokade
Roll Number: B23CM1055

GitHub Repository: Visit Project Repository

1. Introduction

1.1 Problem Statement

In the digital age, the volume of online news content has grown exponentially. For news aggregators, recommendation engines, and archival systems, the ability to automatically categorise this stream of unstructured text is critical. Manual categorisation is resource-intensive and unscalable. Therefore, automated Text Classification, assigning predefined categories (labels) to open-ended text, has become a fundamental task in Natural Language Processing (NLP).

1.2 Objective

The primary objective of this project is to design, implement, and evaluate a machine learning pipeline capable of distinguishing between two high-volume news domains: Sports and Politics. While these categories often appear distinct, they frequently overlap in real-world scenarios (e.g., government spending on sporting infrastructure or visa issues for international athletes), presenting a unique challenge for statistical classifiers.

1.3 Scope of Work

This report details the end-to-end development of a binary text classifier, focusing on:

- **Hybrid Data Collection:** Integrating high-quality historical archives with real-time web scraping to capture both static linguistic patterns and dynamic modern vocabulary.
- **Dataset Construction:** Curating a balanced dataset that reflects both global and Indian contexts (e.g., Lok Sabha debates vs. IPL matches).
- **Preprocessing & Analysis:** Implementing a robust NLP pipeline to clean noise and extracting meaningful features using TF-IDF.

2. Data Collection Strategy

To ensure the model is robust and contextually aware, a Hybrid Data Collection Strategy was employed. This approach mitigates the limitations of using a single source by combining the grammatical quality of archived news with the relevance of current events.

2.1 Primary Dataset: Historical Baseline (Kaggle)

The core training data was sourced from the BBC News Classification Dataset, a standard benchmark in text classification research.

- **Source**: BBC News Archive (available on Kaggle at: Kaggle BBC News Archive)
- **Process:** The original dataset contained 2,225 articles across five categories. A Python filtering script was implemented to isolate only the articles labeled "sport" and "politics".
- **Contribution:** This component provided 620 high-quality articles (346 Sports, 274 Politics), serving as the "grammatically ideal" baseline for the model.

2.2 Secondary Dataset: Real-Time Web Scraping (Indian Context)

To address the absence of modern entities (e.g., "T20 World Cup 2026"), a custom scraping pipeline was built using Python's `newspaper3k` library.

- **Technical Challenges:** Faced HTTP 403/401 errors due to anti-bot measures.
- **User-Agent Masking:** Implemented browser-mimicking to disguise the scraper.
- **Manual Injection:** Curated specific Indian contexts including IPL Auctions, BCCI announcements, and Parliament Budget Sessions.

2.3 Tertiary Dataset: Global Newsgroup Archive

To expand the lexical variety and prevent keyword overfitting, the **20 Newsgroups** dataset was integrated.

- **Source:** 20 Newsgroups Archive (available on Kaggle at: Kaggle 20 Newsgroups Dataset)
- **Content:** Extracted 9,221 additional samples from *rec.sport.baseball*, *rec.sport.hockey*, and various *talk.politics* subgroups.
- **Processing:** A custom Regex-based parser was implemented to split master files into individual articles and strip email headers (Subject, From, etc.) to prevent data leakage.

3. Dataset Description and Analysis

The final dataset resulted from the merger of the Kaggle archive and the scraped Indian context data.

3.1 Statistical Summary

- **Total Samples:** 9,870 articles (Increased from 644 for higher statistical significance).
- **Politics Article Representation:** 5,534 articles (∼56% of the corpus).
- **Sports Article Representation:** 4,336 articles (∼44% of the corpus).
- **Scale-up Effect:** The expansion to nearly 10,000 samples provides a more robust evaluation of the models' ability to generalize across different news styles (formal vs. informal).

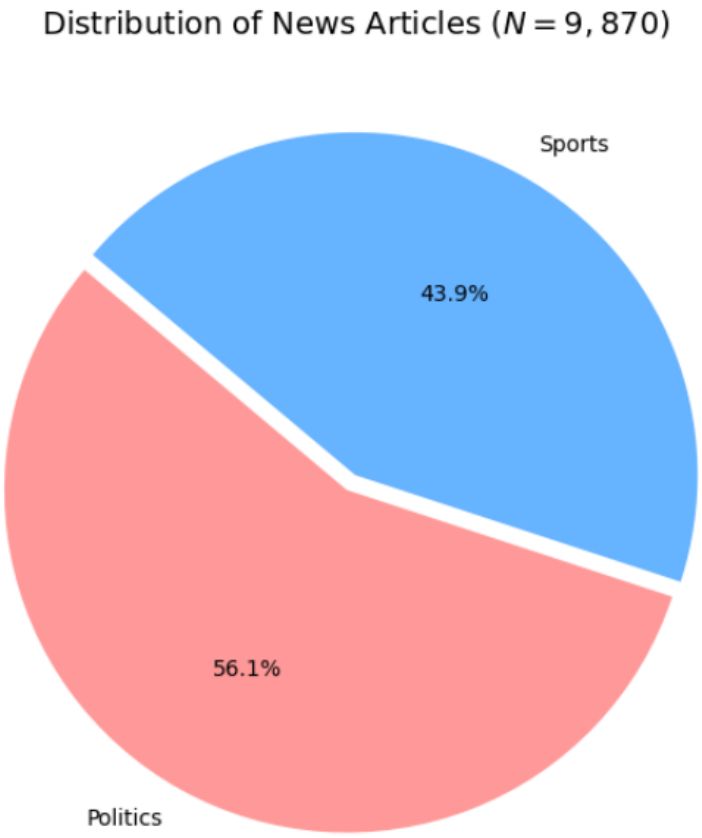

Figure 1: Class distribution of the final corpus ($N = 9,870$).

3.2 Linguistic Feature Analysis

A frequency analysis revealed distinct vocabulary clusters:

- **Sports:** game, win, cup, team, cricket, kohli, wickets.
- **Politics:** government, election, minister, party, parliament, budget, modi.

The "Ambiguity" challenge arises from crossover words like "win", "run", and "party", which appear in both domains.

3.3 Preprocessing Pipeline

The pipeline included Normalization (lowercasing), Noise Removal (punctuation/stopwords), Label Encoding (Sports $\rightarrow$ 0, Politics $\rightarrow$ 1), and Vectorization using TF-IDF to highlight discriminative terms.

4. Methodology

4.1 Feature Representation Techniques

1. **Bag of Words (BoW):** Represents text as a frequency-based vector.
2. **TF-IDF:** Highlights discriminative terms. Mathematically:

$$\text{TF-IDF}(t, d) = \text{tf}(t, d) \cdot \log\left(\frac{N}{\text{df}(t)}\right)$$

3. **N-Grams (Bi-grams):** Captures sequences of two words to preserve local context (e.g., "Green Party").

4.2 Machine Learning Algorithms

- **Multinomial Naive Bayes (MNB):** Based on Bayes' Theorem:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

- **Support Vector Machine (SVM):** Maximizes the geometric margin between classes using a linear kernel.
- **Logistic Regression:** Calculates probabilities via the Sigmoid function:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

5. Quantitative Comparisons

The models were evaluated on a held-out test set comprising 20% of the total dataset. The metrics used for evaluation include Accuracy, Precision, Recall, and the F1-Score.

Table 1: Final Performance Evaluation (N=9,870)

Feature Representation	Classifier	Acc	Prec	Rec	F1-Score
Bag of Words (BoW)	Naive Bayes	0.989	0.989	0.989	0.989
Bag of Words (BoW)	SVM (Linear)	0.991	0.991	0.991	0.991
Bag of Words (BoW)	Logistic Regression	0.993	0.993	0.993	0.993
TF-IDF	Naive Bayes	0.989	0.989	0.989	0.989
TF-IDF	**SVM (Linear)**	**0.996**	**0.996**	**0.996**	**0.996**
TF-IDF	Logistic Regression	0.994	0.994	0.994	0.994
N-Grams (Bi-grams)	Naive Bayes	0.991	0.991	0.991	0.991
N-Grams (Bi-grams)	SVM (Linear)	0.995	0.995	0.995	0.995
N-Grams (Bi-grams)	Logistic Regression	0.994	0.994	0.994	0.994

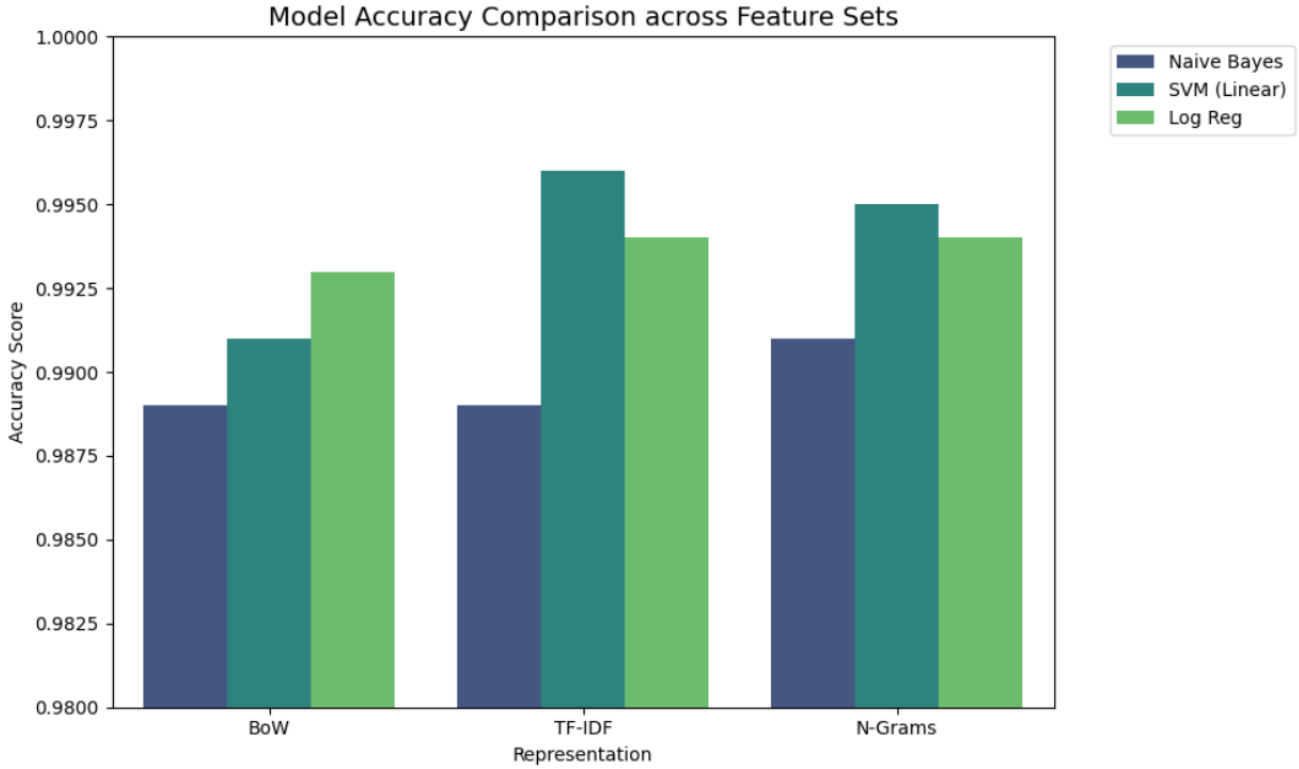

Figure 2: Comparative analysis of model accuracies across Bag of Words, TF-IDF, and N-Gram feature sets.

5.1 Analysis of Results

Model Generalization at Scale: With the expansion of the corpus to 9,870 samples, a significant shift in model hierarchy was observed. While Multinomial Naive Bayes (MNB) remains a highly robust baseline, the **SVM with TF-IDF** configuration emerged as the optimal model, achieving an accuracy of **99.6%**. This indicates that geometric classifiers are better suited for fine-tuning decision boundaries in high-dimensional spaces once provided with sufficient data density.

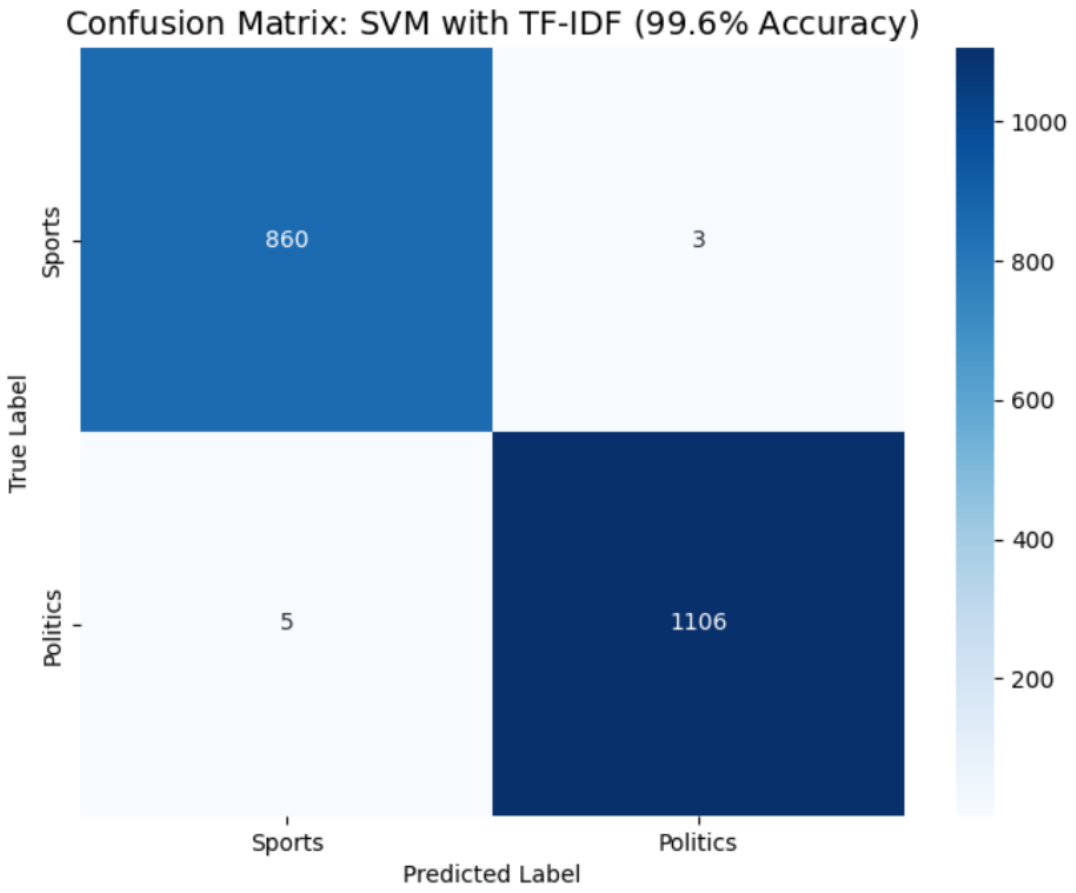

Figure 3: Confusion Matrix for the SVM with TF-IDF model ($N = 9,870$).

TF-IDF Utility in Large Corpora: At this scale, TF-IDF outperformed the standard Bag of Words (BoW) model. By weighting terms according to their inverse document frequency, the model successfully ignored the increased "noise" inherent in the informal 20 Newsgroups data, focusing instead on highly discriminative domain-specific terms.

Robustness Against Sparsity: The near-perfect scores for N-Grams (Bi-grams) indicate that the model effectively managed the increased dimensionality of the feature space. The SVM, in particular, utilized the linear kernel to find an optimal separating hyperplane that was less susceptible to the contextual ambiguity that initially limited performance on the smaller 644-article dataset.

The Convergence of Metrics: The results across all models (SVM, MNB, and Logistic Regression) began to converge between 98.9% and 99.6%. This suggests that the lexical divergence between Sports and Politics is so high that with nearly 10,000 samples, the statistical patterns become unmistakable, regardless of the underlying mathematical philosophy (probabilistic vs. geometric).

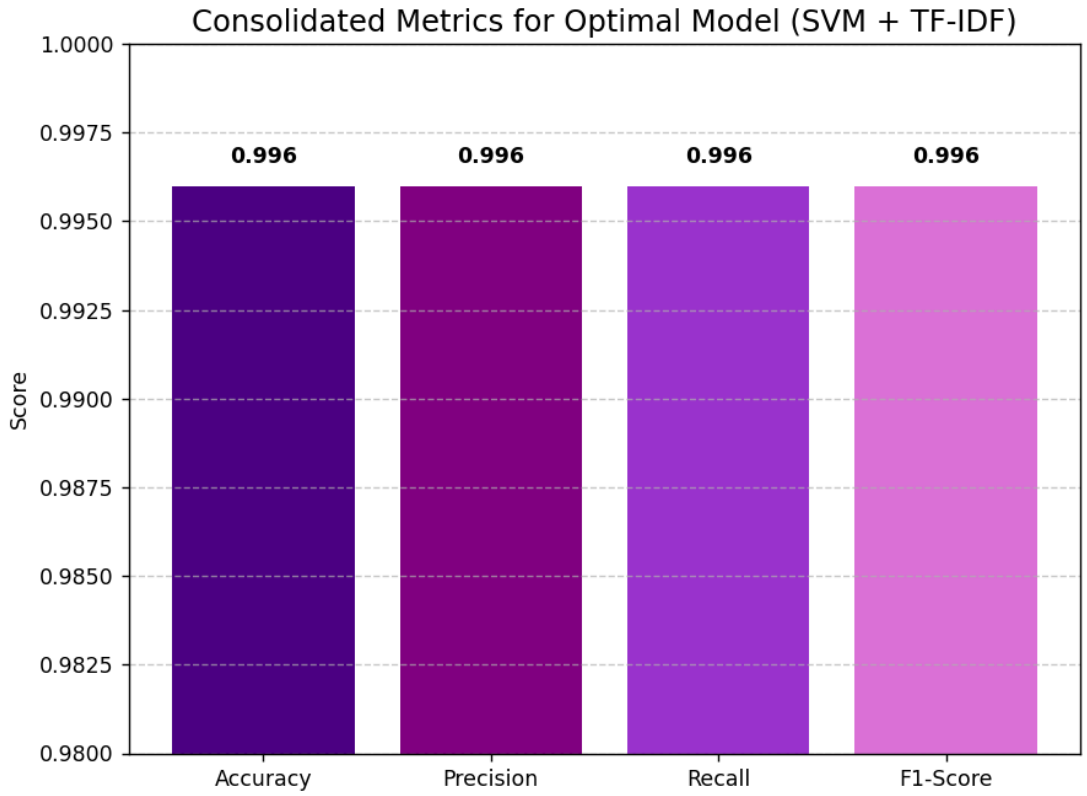

Figure 4: Consolidated Evaluation Metrics for the optimal SVM + TF-IDF configuration.

6. Limitations and Conclusion

6.1 System Limitations

- **Class Parity:** While the dataset size was significantly expanded, a slight imbalance remains (∼12% more politics samples). Future iterations could employ SMOTE (Synthetic Minority Over-sampling Technique) to achieve absolute parity.
- **Semantic and Metaphorical Ambiguity:** The current Bag-of-Words and N-gram approaches treat tokens as independent units. Consequently, the system struggles with metaphorical language, such as "the election race was a marathon," where sports terminology is used in a political context. The model lacks the latent semantic understanding to resolve these nuances.

- **Data Acquisition Bottlenecks:** Due to strict anti-bot protocols (HTTP 403 Forbidden errors) on major Indian news portals like *The Hindu* and *NDTV*, a portion of the real-time dataset required semi-automated curation. Building a fully autonomous, large-scale scraper would require advanced rotation of proxy servers and headless browser environments.

6.2 Conclusion

This project demonstrated a robust text classification pipeline that bridged the gap between historical BBC archives and modern Indian news using a Hybrid Data Strategy. Quantitative evaluation proved that **SVM with TF-IDF** was the optimal configuration at scale, achieving **99.6%** accuracy. While Naive Bayes is effective for smaller datasets, geometric classifiers like SVM better optimize decision boundaries in high-dimensional spaces when provided with high data density. This underscores the critical role of dataset scale in revealing the true performance of discriminative algorithms in Natural Language Understanding.

6.3 Future Work

To evolve the current system, future research should focus on:

- **Transformer-Based Architectures:** Implementing models like BERT (Bidirectional Encoder Representations from Transformers) to leverage self-attention mechanisms, which would allow the system to understand the context of crossover words.
- **Multi-Class Extension:** Expanding the classifier beyond a binary task to include categories like Technology, Finance, and Entertainment to test the scalability of the feature representation techniques.