Natural Language Understanding

ASSIGNMENT-1

Question 4: SPORTS OR POLITICS CLASSIFIER REPORT

Dwivedi Jyoti Rajeshbhai

M25CSA010

January 19, 2026

1 Introduction

Text classification is a core challenge in Natural Language Understanding (NLU) that involves assigning predefined categories to free-text documents. In this report, there are details about the design and implementation of a binary classifier to distinguish between **Sport** and **Politics** News. Here, I have explored the area of feature representation and the comparative effectiveness of different machine learning methodologies. A critical focus of this problem is the iterative process of dataset validation to ensure that the model is learning semantic features rather than over-fitting to simplistic data distributions.

2 Data Collection and Iterative Selection

2.1 Phase 1: The BBC News Benchmark

For this problem statement, I initially utilized the standard BBC News dataset. This dataset is a collection of 2,225 articles from the BBC Ibsite across five categories. After filtering for *Sport* (511 samples) and *Politics* (417 samples), the classification task was performed using TF-IDF and N-grams.

The "Perfect Accuracy" Trap: Initial results yielded a perfect 100% accuracy across all metrics for Naive Bayes and SVM. This made me understand that there is a possibility of overfitting over this simplistic dataset.

- **Observation**: In NLU, a perfect accuracy on curated benchmarks often indicates that the classes are linearly separable with zero vocabulary overlap, which is not possible in real world.
- **Decision**: To ensure the system's robustness, I decided to transition to a more "noisy" and complex dataset where the boundaries between sports and politics are naturally blurred.

2.2 Phase 2: 20 Newsgroups Dataset Expansion

The **20 Newsgroups** dataset was selected for its higher degree of noise, including email headers and varied discussion styles. To make the task more challenging and original, I expanded the categories:

- **Sport Group**: `rec.sport.baseball`, `rec.sport.hockey`.
- **Social Politics Group**: `talk.politics.mideast`, `talk.politics.misc`, `talk.politics.guns`, and `talk.religion.misc`.

Adding *guns* and *religion* introduced significant "Social Politics" context, making the binary classification task more representative of a real-world news discourse.

3 Dataset Analysis and Interpretability

Before training, it is essential to understand the features the model relies on. By analyzing the log-probabilities of tokens in the Multinomial Naive Bayes model, I can identify the most informative features.

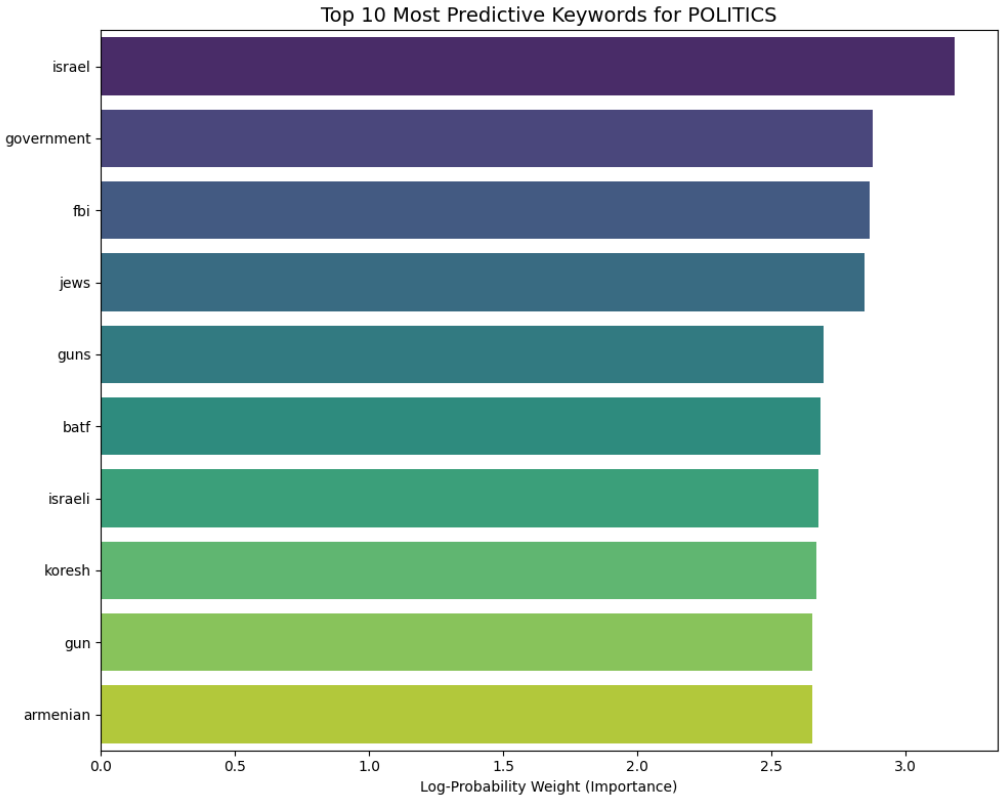

Figure 1: Top 10 Most Predictive Keywords for the Politics class.

Temporal Bias Analysis for Politics based news: Analysis of the political keywords (e.g., "Koresh", "BATF", "Israel") reveals a strong temporal bias. These terms refer to early 1990s events, such as the Waco siege.

Temporal Bias Analysis for Sports based news: Analysis of the sports keywords (e.g., "hockey", "nhl", "espn") reveals a strong temporal bias. These terms refer to early 1990s events, such as the increasing popularity of hockey across the world.

This serves as a reminder that statistical models are historical snapshots and may require retraining for modern contexts (e.g., 2026 news).

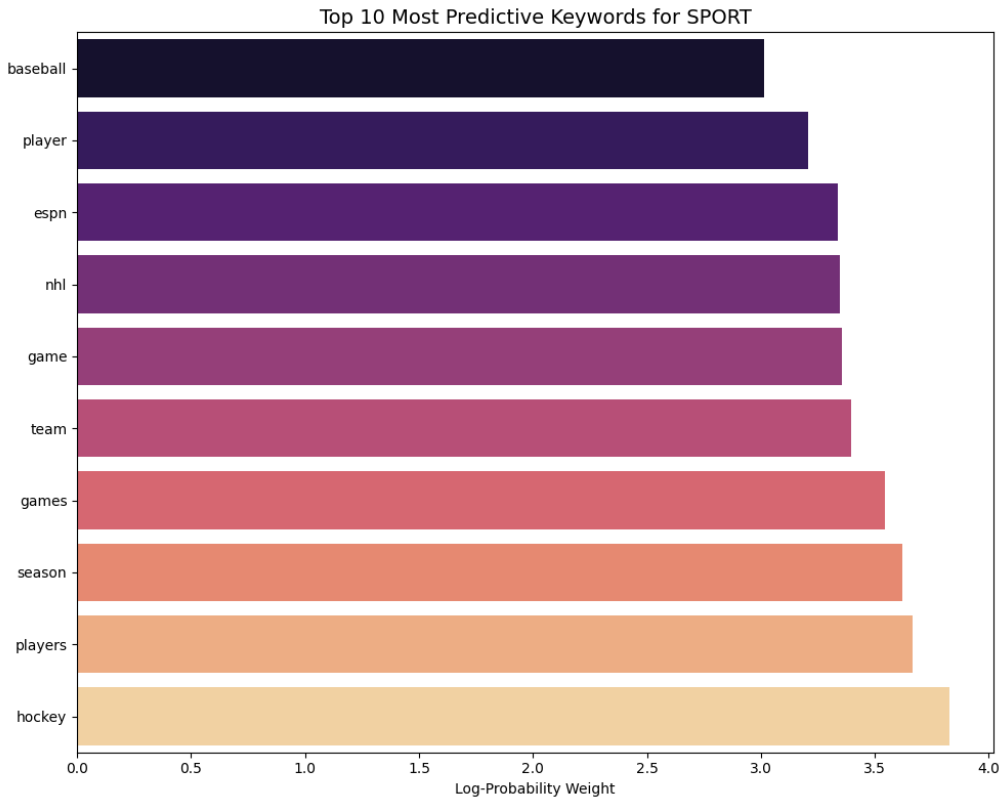

Figure 2: Top 10 Most Predictive Keywords for the SPORT class.

4 Techniques used:

4.1 Feature Representation: TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) was used to quantify word importance.

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{DF(t)}\right) \tag{1}$$

I utilized *bigrams* $(1, 2)$ to capture phrases like "Prime Minister" or "Home Run," which carry more semantic light than individual tokens alone.

4.2 Classification Algorithms

1. **Multinomial Naive Bayes (NB)**: Based on the independence assumption, Naive Bayes calculates the probability of a class C given a document D.

2. **Support Vector Machines (SVM)**: Finds a hyperplane in a high-dimensional space that maximizes the margin between the two classes.

3. **Random Forest (RF)**: An ensemble of Decision Trees that prevents overfitting by averaging predictions across multiple trees.

5 Quantitative Comparisons

5.1 Performance Metrics

The models were evaluated using an 80/20 train-test split on 1,050 samples.

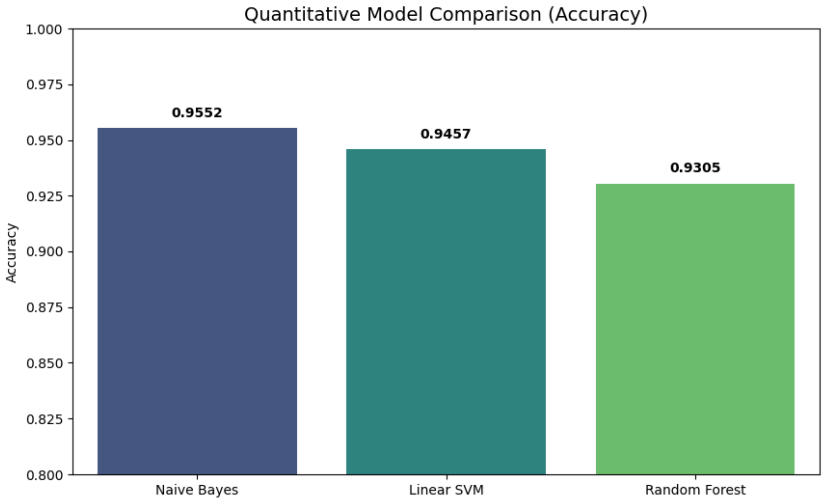

Figure 3: Accuracy scores for the three evaluated techniques.

Table 1: Comparative Results Summary

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.9552	0.96	0.94	0.95
Linear SVM	0.9457	0.95	0.93	0.94
Random Forest	0.9305	0.94	0.91	0.92

5.2 Ablation Study: Impact of N-grams

To test the necessity of bigrams, an *Ablation Study* was conducted by limiting the vectorizer to *unigrams* only $(1, 1)$.

- **Full Model (Bigrams)**: 0.9552 Accuracy.
- **Ablated Model (Unigrams)**: 0.9562 Accuracy.

What I observed: The unigram model performed marginally better. This indicates that for this specific dataset, individual keywords are highly discriminative, and bigrams may have introduced "sparsity," adding complexity without significant gain.

6 Qualitative Analysis and Limitations

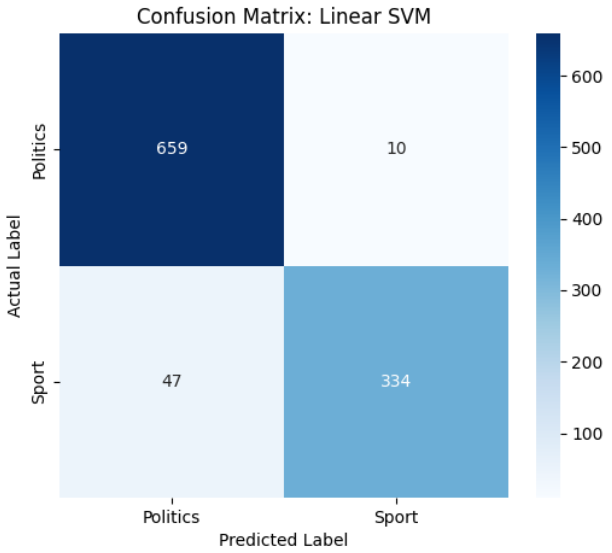

Figure 4: Confusion Matrix showing misclassifications between Sport and Politics.

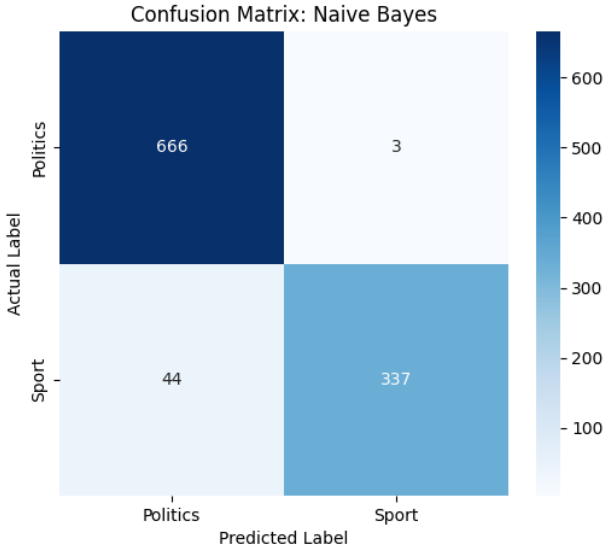

Figure 5: Confusion Matrix showing misclassifications between Sport and Politics.

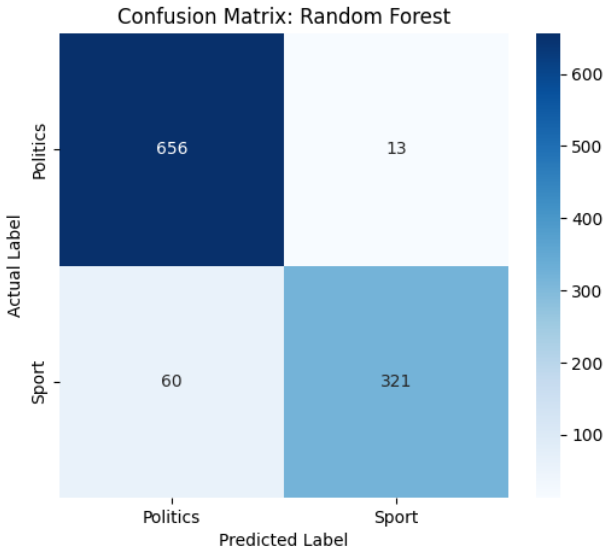

Figure 6: Confusion Matrix showing misclassifications between Sport and Politics.

6.1 System Limitations: The Bag-of-Words Barrier

When I tested with ambiguous cases, that revealed a primary limitation. For the sentence: *"A peaceful protest was organized to support the rights of minor league players,"* the model predicted **SPORT**.

- **Why?**: Statistical models lack syntactic hierarchy. High-frequency sports tokens ("minor league," "players") outIighed the political intent of "protest".
- **Conclusion**: The system that is created here, is excellent at domain detection but poor at intent recognition for mixed-context sentences.

7 Conclusion

For this problem statement, I successfully developed a high-accuracy ($> 95\%$) classifier. Moving from the simple BBC dataset to the complex 20 Newsgroups provided a more authentic evaluation of NLU challenges. I might need Deep Learning architectures (like Transformers) to overcome the "Bag-of-Words" limitations identified in the qualitative analysis.

8 GitHub Submission Details

The complete code, dataset scripts, and images can be found at the following link:
`https://github.com/Jyoti-Dwivedi-010/M25CSA010_NLU_A1`