CSL 7640: Natural Language Understanding

Problem 4: Comparative Study of News Classification using TF-IDF and Linear Models

Author:

Bhagwan Arsewad

Roll Number: B22AI010

Department: AI and Data Science

Institution: Indian Institute of Technology Jodhpur

Date of Submission: February 15, 2026

1 Dataset Acquisition and Exploratory Analysis

The experimental framework utilized a corpus of **928 news documents** extracted from the BBC news archive. The primary goal was to distinguish between the domains of **Politics** and **Sports**.

1.1 Document Distribution and Balance

The dataset follows a class-labeled directory structure. A balanced distribution is essential for avoiding bias in probabilistic models like Naive Bayes:

- **Politics (417 files):** Focuses on government policy, electoral news, and parliamentary debates.
- **Sport (511 files):** Focuses on competitive match reports, tournament statistics, and player profiles.

1.2 Linguistic Feature Exploration

A critical step in my analysis involved identifying the words most responsible for class separation. The following terms showed the highest TF-IDF weights in my terminal execution:

- **Politics Signature:** *mr, labour, election, blair, government, party, brown, minister.*
- **Sport Signature:** *game, england, win, world, cup, play, players, match.*

2 Technical Methodology and Feature Engineering

The conversion of unstructured text into a structured numerical format is the cornerstone of this classification pipeline.

2.1 Vectorization: TF-IDF Framework

I implemented **TF-IDF** (Term Frequency-Inverse Document Frequency) to transform text into a high-dimensional sparse matrix. Unlike simple Bag-of-Words, **TF-IDF** weights a term t in document d as:

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

2.2 The Impact of N-Grams

By incorporating **Bigrams** $(n = 1, 2)$, the model captures unified semantic features. For example, the term "World" and "Cup" might appear independently in many contexts, but the bigram "World Cup" is an extremely strong indicator for the Sports category.

2.3 Stop Word Removal and Noise Reduction

To improve the signal-to-noise ratio, I utilized a standard English stop-word list. This ensured that common tokens like "the", "is", and "and" did not dilute the TF-IDF weights of domain-specific terms like "Labour" or "Arsenal".

3 Supervised Learning: Algorithm Benchmarking

I benchmarked three distinct supervised learning techniques to evaluate their robustness in a linearly separable feature space.

3.1 Multinomial Naive Bayes (MNB)

MNB is a probabilistic model based on Bayes' theorem. It assumes conditional independence between features given the class. Despite this "naive" assumption, it is exceptionally fast and performs well in high-dimensional text classification.

3.2 Logistic Regression (LR)

LR is a linear classifier modeling the probability of class membership via the sigmoid function. It is highly interpretable, as the learned coefficients directly indicate the importance of specific words in the decision-making process.

3.3 Support Vector Machines (SVM)

I utilized a Linear SVM to identify the maximum-margin hyperplane $w^T x + b = 0$. SVMs are particularly effective in text classification because they can handle high-dimensional sparse data without significant overfitting.

4 Quantitative Results and Analysis

The models were evaluated using an 80/20 train-test split. The metrics used include Accuracy, Precision, Recall, and the F1-Score.

```
PS C:\Users\arsew\OneDrive\Desktop\8 Sem\NLU\B22AI010_A1> python B22AI010_prob4.py
Step 1: Loading 928 text documents...
Step 2: Transforming text into TF-IDF vectors...

Step 3: Evaluating Models...
------------------------------------------
| Multinomial Naive Bayes   | Accuracy: 1.0000 |
| Logistic Regression       | Accuracy: 1.0000 |
| Linear SVM                | Accuracy: 1.0000 |

--- Top Discriminative Bigrams per Category ---
POLITICS: mr, said, labour, election, blair, government, party, brown, people, minister
SPORT: said, game, england, win, year, world, cup, play, players, match
PS C:\Users\arsew\OneDrive\Desktop\8 Sem\NLU\B22AI010_A1>
```

Figure 1: Captured Terminal Results showing 100% metric performance.

Technique	Accuracy	Precision	Recall	F1-Score
Naive Bayes	1.0000	1.00	1.00	1.00
Logistic Regression	1.0000	1.00	1.00	1.00
Linear SVM	1.0000	1.00	1.00	1.00

Table 1: Consolidated Performance Metrics.

4.1 Linear Separability Discussion

The **1.0000 Accuracy** suggests that the dataset is "Linearly Separable". In high-dimensional spaces, it is common for distinct categories like Sports and Politics to have non-overlapping vocabularies, allowing models to achieve perfect separation on the test set.

5 Critical Limitations and Future Scope

While the results demonstrate high efficacy, several constraints exist for real-world deployment.

5.1 Key Limitations

- **Contextual Ambiguity:** Simple keyword models cannot distinguish crossover articles, such as a report on "The politics of the 2024 Olympic stadium construction".
- **Vocabulary Drift:** The model lacks the ability to understand new terminology (e.g., a newly formed political party) without full retraining.
- **Informal Text Handling:** The current system is trained on formal news; accuracy would likely degrade on noisy social media text or slang.

5.2 Future Work

Future iterations could explore deep learning architectures such as LSTMs or Transformers (BERT) to capture long-range dependencies and semantic context beyond simple n-grams. Additionally, active learning could be implemented to handle new categories with minimal labeled data.

6 GitHub Submission

The full source code and logs are hosted at:
https://github.com/bhagwan388/NLU-B22AI010_A1

7 Conclusion

This study proves that simple linear models, when paired with N-gram TF-IDF representations, provide a highly reliable baseline for domain-specific news classification.