1). Dataset Details

- **Shape:** (2225, 2)
- **Categories:** sport, business, politics, tech, entertainment
- **Distribution:**
  - sport → 511
  - business → 510
  - politics → 417
  - tech → 401
  - entertainment → 386

---

2). Preprocessing Check

Example of processed text:

Original: "tv future in the hands of viewers with home theatre system..."

Processed: "tv future hand viewer home theatre system plas..."

---

3). Model Results

1. Naive Bayes

- **Accuracy:** 98.42%
- **Performance:**
  - Business → Precision 0.97, Recall 0.97, F1 0.97
  - Entertainment → Precision 0.99, Recall 0.99, F1 0.99
  - Politics → Precision 0.97, Recall 0.99, F1 0.98
  - Sport → Precision 1.00, Recall 1.00, F1 1.00
  - Tech → Precision 1.00, Recall 0.97, F1 0.99

---

2. Logistic Regression

- **Accuracy:** 98.42%
- **Performance:**
  - Business → Precision 0.97, Recall 0.98, F1 0.98
  - Entertainment → Precision 0.97, Recall 1.00, F1 0.99
  - Politics → Precision 0.98, Recall 0.99, F1 0.98

- Sport → Precision 1.00, Recall 1.00, F1 1.00
- Tech → Precision 1.00, Recall 0.95, F1 0.97

---

3. Support Vector Machine (SVM)

- **Accuracy:** 98.42%
- **Performance:**

business → precision 0.98 , recall 0.97 , f1 0.98

entertainment → precision 0.97 , recall 1.00, f1 0.99

politics → precision 0.98 , recall 0.99 , f1 0.98

sport → precision 1.00 , recall 1.00 , f1 1.00

tech → precision 0.99 , recall 0.96 , f1 0.97

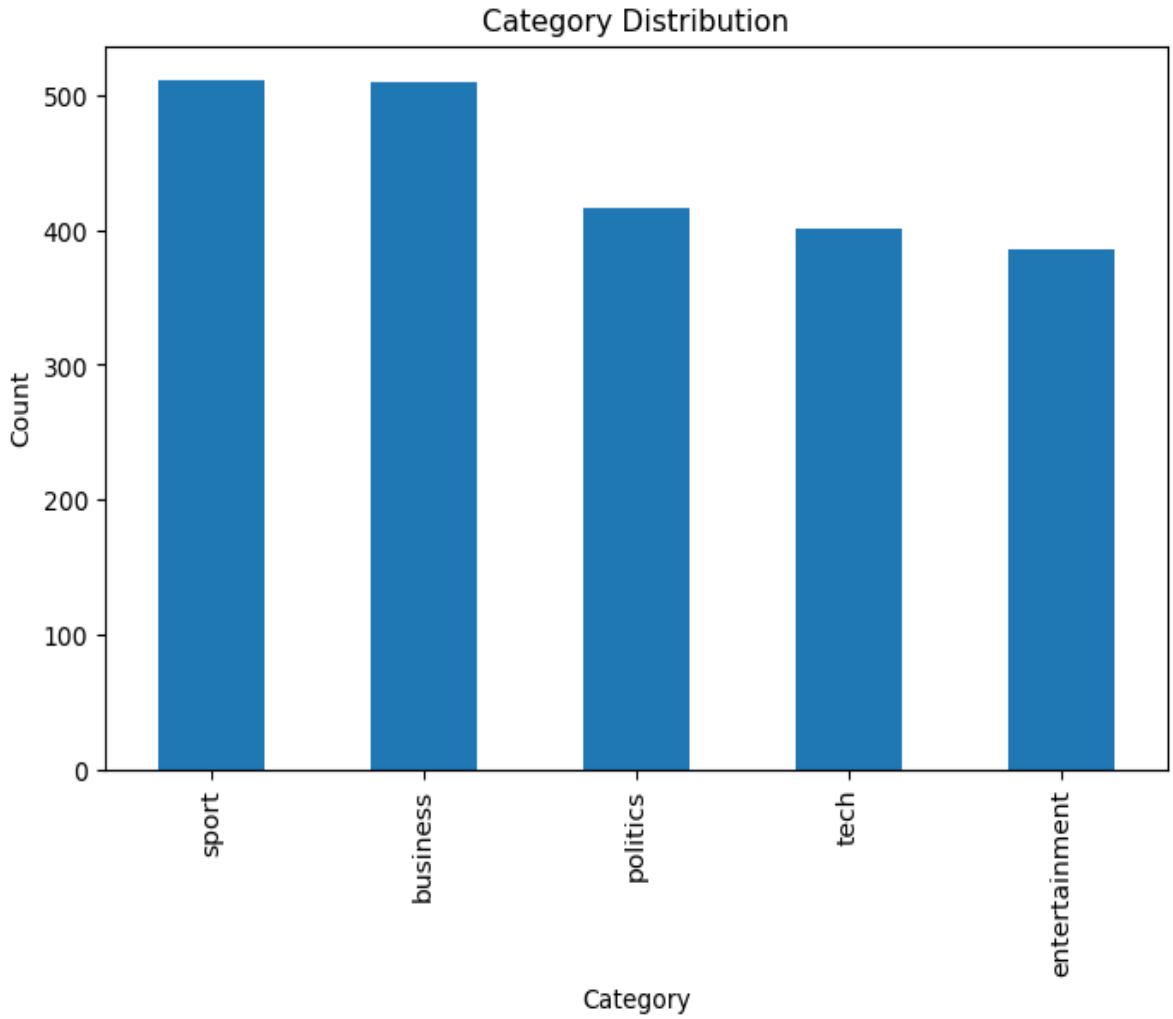
---

4). Category Distribution

The bar chart represents the **distribution of documents across the five categories** in the dataset: *sport, business, politics, tech,* and *entertainment*.

- **Sport (511)** and **Business (510)** have the highest number of documents.
- **Politics (417)** and **Tech (401)** are moderately represented.
- **Entertainment (386)** has the fewest documents.

👉 This distribution is **fairly balanced** across categories, which is beneficial for training classification models since no single class dominates the dataset.



5). Confusion Matrix – Naive Bayes

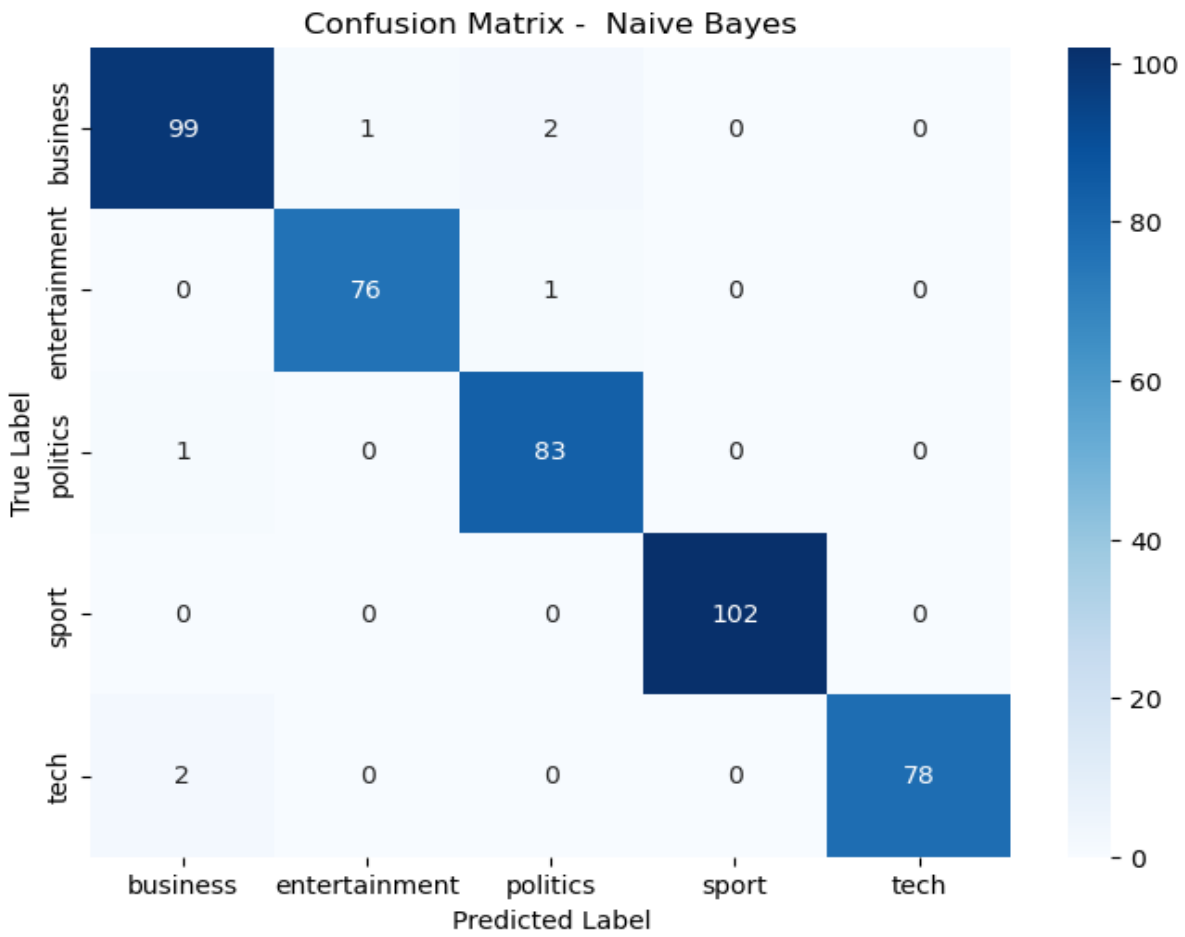
This confusion matrix shows the performance of the Naive Bayes classifier across the five categories (*business, entertainment, politics, sport, tech*).

- **Diagonal values** represent correctly classified documents.
- **Off-diagonal values** represent misclassifications.

🔍 **Observations:**

- **Business:** 99 correctly predicted, with only 3 misclassifications (1 as entertainment, 2 as politics).
- **Entertainment:** 76 correctly predicted, only 1 misclassified as politics.
- **Politics:** 83 correctly predicted, just 1 misclassified as business.
- **Sport:** Perfect classification with 102 correct predictions (no misclassifications).
- **Tech:** 78 correctly predicted, only 2 misclassified as business.

✅ **Conclusion:**
The model achieves **very high accuracy** with minimal misclassifications. Most errors occur between **business and politics** or **business and tech**, which is expected since these categories sometimes share overlapping vocabulary.



6). Confusion Matrix – Logistic Regression

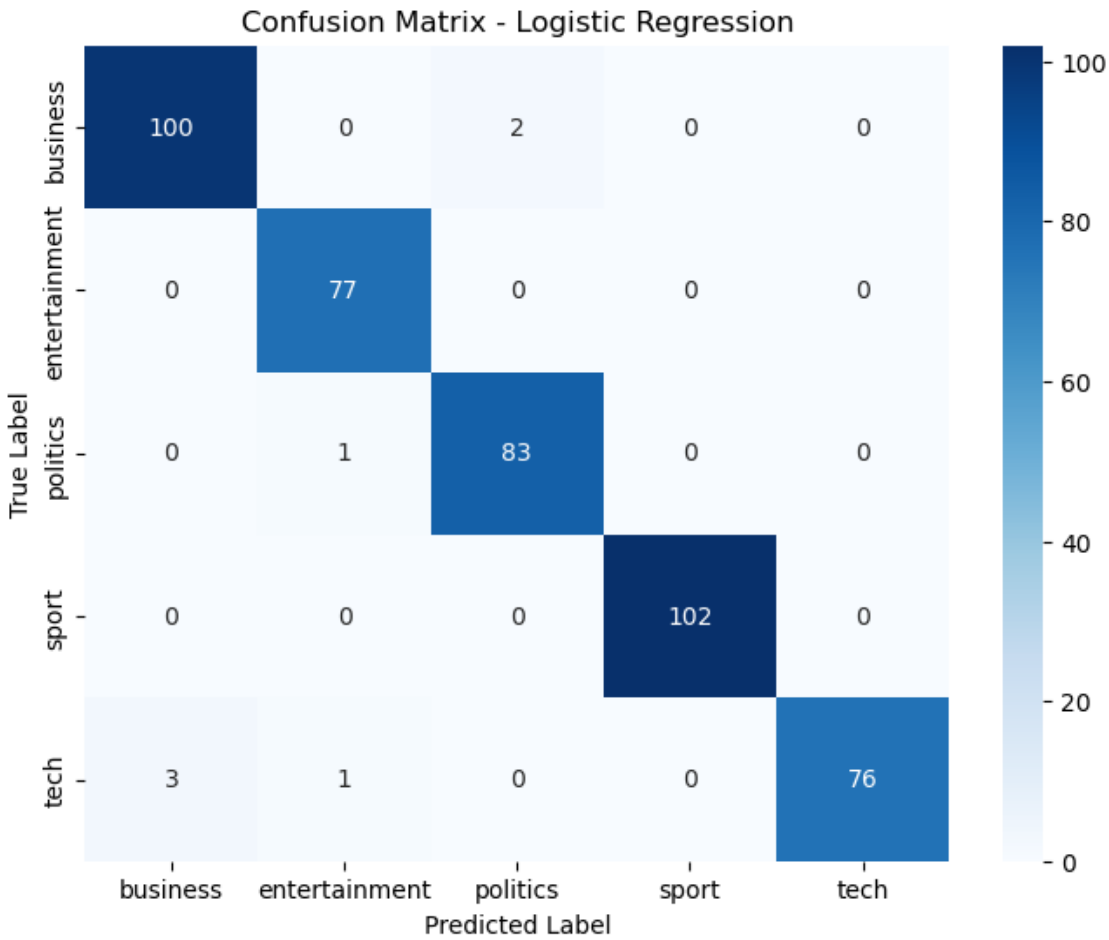
This confusion matrix shows how Logistic Regression classified documents across the five categories (*business, entertainment, politics, sport, tech*).

- **Diagonal values** → correctly predicted documents.
- **Off-diagonal values** → misclassified documents.

🔍 **Observations:**

- **Business:** 100 correctly predicted, only 2 misclassified as politics.
- **Entertainment:** 77 correctly predicted, no major misclassifications.
- **Politics:** 83 correctly predicted, 1 misclassified as entertainment.
- **Sport:** Perfect classification with all 102 correctly predicted (no errors).
- **Tech:** 76 correctly predicted, with 3 misclassified as business and 1 as entertainment.

✅ **Conclusion:**
The Logistic Regression model achieves **excellent accuracy** with very few misclassifications. Most errors occur between **tech and business** categories, likely due to overlap in business-technology terminology.

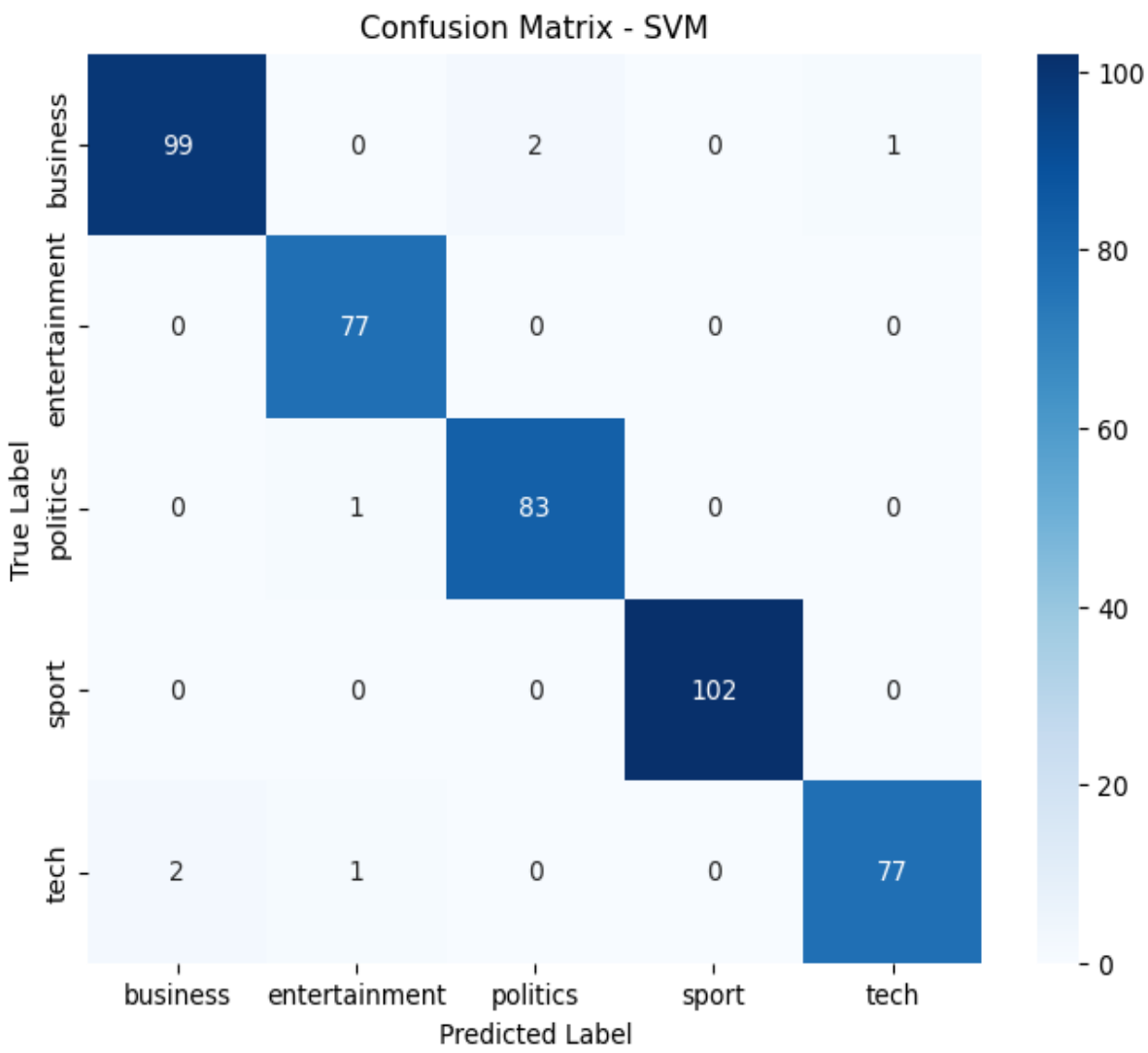


7). **Confusion Matrix – Support Vector Machine (SVM)**

This confusion matrix illustrates how the SVM model classified documents across the five categories (*business, entertainment, politics, sport, tech*).

- **Diagonal values** → correct predictions.
- **Off-diagonal values** → errors.

🔍 **Observations:**

- **Business:** 99 correctly predicted, with 2 misclassified as politics and 1 as tech.
- **Entertainment:** 77 correctly predicted, no major errors.
- **Politics:** 83 correctly predicted, 1 misclassified as entertainment.
- **Sport:** Perfect classification with 102 correctly predicted (no misclassifications).
- **Tech:** 77 correctly predicted, with 2 misclassified as business and 1 as entertainment.

✅ **Conclusion:**
The SVM model demonstrates **very high performance**, with only a handful of misclassifications. Errors are minimal and mostly occur between **business and tech**—a common overlap due to shared vocabulary.