Performance Evaluation Report: News Article Classifier AI

1. Introduction & Implementation Summary

This report details the performance analysis of the new AI model designed to automatically classify news articles into one of three categories: 'Technology', 'Sports', or 'Business'. The evaluation was conducted on a sample dataset of 20 articles, comparing the model's predicted categories against the manually verified true categories.

The analysis was implemented in Python using the following standard data science libraries:

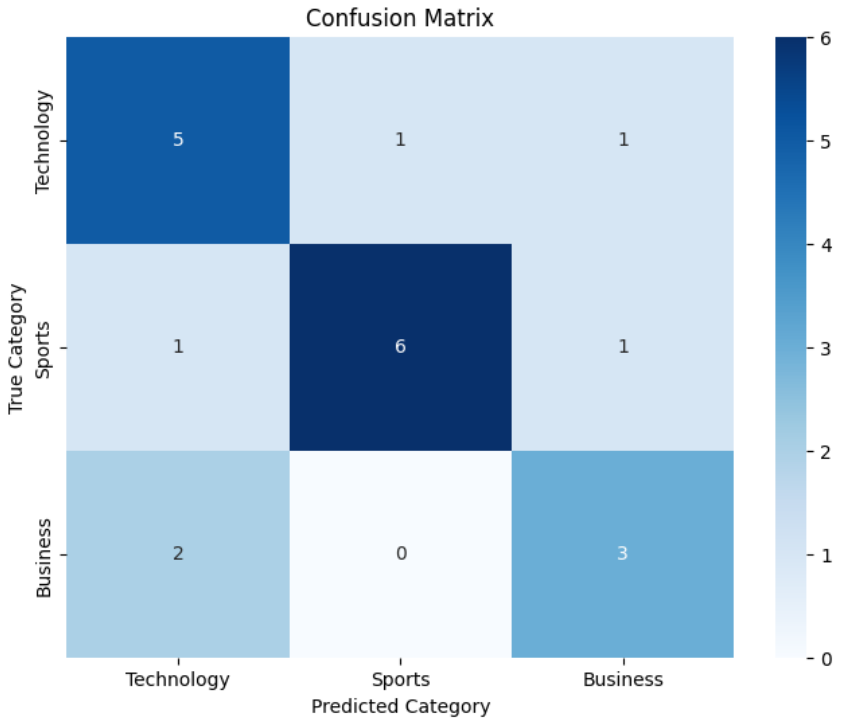
- **Pandas:** For managing the dataset of true and predicted labels.
- **Scikit-learn:** For the core analysis, including the calculation of the confusion matrix, overall accuracy, and a detailed classification report with per-class metrics.
- **Seaborn & Matplotlib:** For generating a clear and intuitive heatmap visualization of the confusion matrix.

2. The Multi-Class Confusion Matrix

A confusion matrix is a table that summarizes the performance of a classification model. For a multi-class problem like this one, it shows exactly how the predictions for each category line up against the true categories, revealing where the model is succeeding and where it is getting "confused."

Our Model's Confusion Matrix

The 3x3 matrix below was generated from the 20-article test set. The rows represent the actual category of an article, and the columns represent the category the model predicted.



3. Analysis of Performance Metrics

Overall Accuracy: 70.0%

- **Explanation:** This is the single score representing the percentage of all articles that the model categorized correctly. An accuracy of 70% indicates a moderate level of performance.
- **Calculation:** (5 + 6 + 3) / 20

Per-Class Metrics

- **Precision:** Answers the question, "When the model predicts a certain category, how often is it correct?"
- **Recall:** Answers the question, "Of all the articles that truly belong to a category, how many did the model correctly identify?"
- **F1-Score:** The harmonic mean of Precision and Recall, providing a single score that balances both for a specific class.

Category	Precision	Recall	F1-Score
Technology	62.0%	71.0%	67.0%
Sports	86.0%	75.0%	80.0%
Business	60.0%	60.0%	60.0%

Macro-Averaged Metrics

These are the simple arithmetic average of the per-class metrics, treating each category with equal importance.

- **Macro-Averaged Precision:** 69.0%
- **Macro-Averaged Recall:** 69.0%
- **Macro-Averaged F1-Score:** 69.0%

4. Short Analysis and Recommendation

Based on the new metrics, the model's overall performance is modest at 70% accuracy. The model handles the **'Sports'** category the best, demonstrated by its high F1-score (80%) and strong precision (86%). Its biggest problem is with the **'Business'** category, which has the lowest F1-score (60%) and is frequently misclassified as 'Technology'.

Recommendation: The model is not yet reliable enough for unmonitored deployment. Its primary weakness is its inability to consistently distinguish 'Business' articles, especially from 'Technology'. We recommend targeted retraining with more examples of 'Business' and 'Technology' articles to help the model learn the distinguishing features between them.