**Amharic Text Classification Using Logistic Regression**

1. Objective

The goal of this project is to classify Amharic news headlines into predefined categories such as Politics, Sport, Technology, and Health using a supervised machine learning approach. The model is trained to automatically label headlines based on patterns learned from a labeled dataset.

2. Tools and Libraries

beautifulsoup4==4.13.4

joblib==1.4.2

matplotlib==3.10.3

numpy==2.2.5

pandas==2.2.3

Requests==2.32.3

scikit\_learn==1.6.1

seaborn==0.13.2

3. Dataset Collection and Preprocessing

The dataset was scraped from BBC Amharic using BeautifulSoup. Headlines from four categories (Politics, Health, Technology, Sport) were collected across multiple pages to ensure balance. A total of 466 samples were cleaned and normalized. Punctuation was removed and whitespace trimmed. To balance the dataset, each class was undersampled to 112 samples.

4. Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency) was used to convert text to numerical vectors. Bigrams were included to capture word pairings, and up to 8000 features were extracted.

5. Model Training and Evaluation

Logistic Regression was used as the classifier, trained using 80% of the balanced dataset. The model was evaluated using accuracy, precision, recall, and F1-score. The following results were obtained:

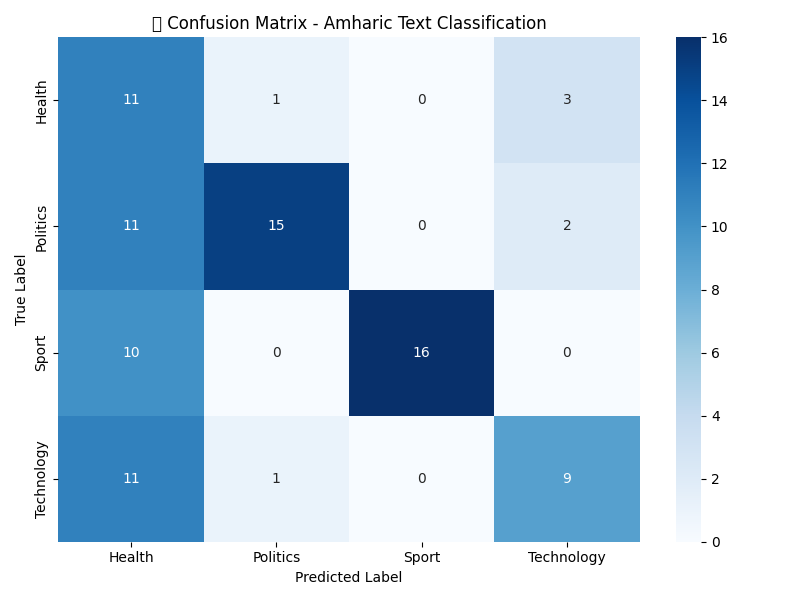
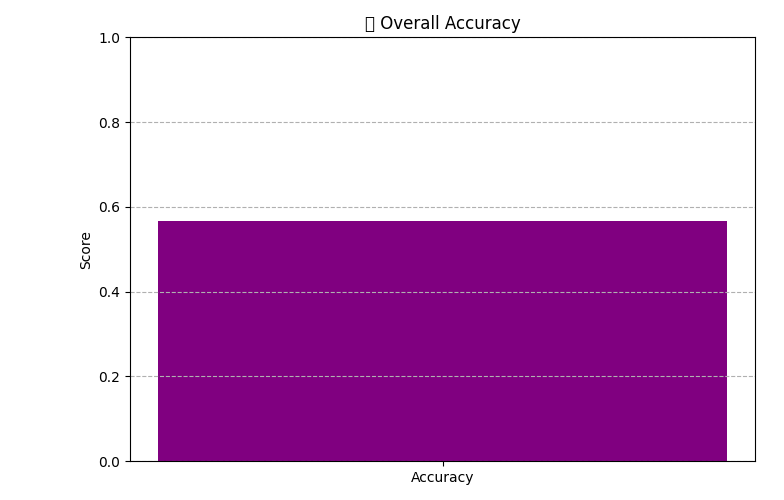
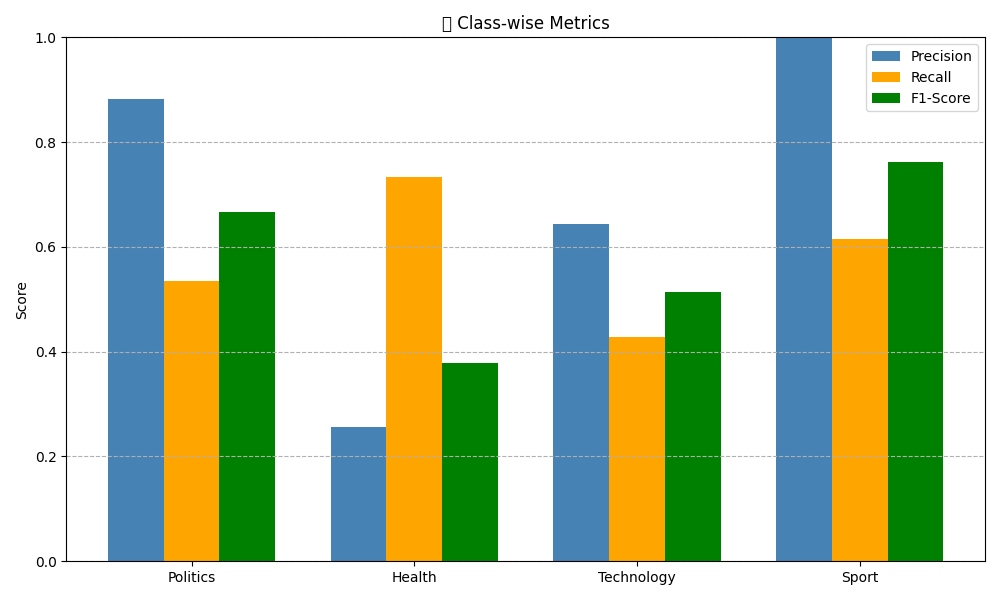
Accuracy: 57%  
Macro F1-score: 58%  
  
Class-wise F1 scores:  
- Health: 0.38  
- Politics: 0.67  
- Sport: 0.76  
- Technology: 0.51

6. Confusion Matrix (Simplified)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Health | Politics | Sport | Technology |
| Health | 11 | 1 | 0 | 3 |
| Politics | 11 | 15 | 0 | 2 |
| Sport | 10 | 0 | 16 | 0 |
| Technology | 11 | 1 | 0 | 9 |

7. Conclusion and Future Work

The model shows reasonable performance considering the dataset size. Sport and Politics were classified most accurately. Health and Technology need more distinct data or preprocessing. Future improvements include:  
- Collecting more balanced and diverse samples  
- Using word embeddings (e.g., fastText or BERT for Amharic)  
- Exploring advanced deep learning models like LSTM or transformers



Note About measurements.

**First: What’s a "Good Prediction"?**

When your model predicts:

* 🎯 Correctly → it’s a "True Positive" or "True Negative"
* ❌ Incorrectly → it’s a "False Positive" or "False Negative"

These terms form the basis of all evaluation metrics.

**✅ 1. Accuracy — "How many did I get right?"**

**📌 Formula:**

Accuracy = (Correct predictions) / (Total predictions)

= (TP + TN) / (TP + FP + FN + TN)

**🧠 Example:**

Out of 100 Amharic headlines:

* 57 were predicted correctly
* So, accuracy = 57 / 100 = **57%**

**✅ 2. Precision — "When I predict a category, how often am I right?"**

**📌 Formula:**

Precision = TP / (TP + FP)

* TP = True Positive
* FP = False Positive

**🧠 Example:**

Out of 20 headlines predicted as **“Sport”**:

* 15 were actually Sport
* 5 were mistakes (maybe they were Politics or Health)

So precision = 15 / (15 + 5) = **0.75 (75%)**

🗣️ *“When I say it’s Sport, I’m right 75% of the time.”*

**✅ 3. Recall — "How many real cases did I catch?"**

**📌 Formula:**

Recall = TP / (TP + FN)

* FN = False Negatives (missed positives)

**🧠 Example:**

There are 30 real “Health” headlines.

* Your model predicted 18 of them correctly
* Missed 12 (classified them as something else)

So recall = 18 / (18 + 12) = **0.6 (60%)**

🗣️ *“Out of all actual Health news, I caught 60%.”*

**✅ 4. F1 Score — "Balance between precision and recall"**

**📌 Formula:**

F1 Score = 2 \* (Precision \* Recall) / (Precision + Recall)

* It’s the **harmonic mean** of precision and recall
* Gives a **single number** that balances both

🧠 If precision is high but recall is low (or vice versa), F1 helps you see that balance.

**📊 In Your Classification Report:**

You’ll usually see something like:

precision recall f1-score support

Health 0.72 0.45 0.55 20

Politics 0.88 0.90 0.89 30

| **Metric** | **Meaning** |
| --- | --- |
| Precision | Of the predicted "Health", how many were correct? |
| Recall | Of the actual "Health" items, how many did you catch? |
| F1 Score | A balanced score between precision & recall |
| Support | Number of true samples in the dataset for that class |
|  |  |

**🧠 Summary Table**

| **Metric** | **Focus** | **Question It Answers** |
| --- | --- | --- |
| Accuracy | Overall correctness | How many total predictions were right? |
| Precision | Prediction quality | When I predict class A, how often am I right? |
| Recall | Sensitivity | Did I find all the real examples of class A? |
| F1 Score | Balance of P & R | How good is the tradeoff between P and R? |