Model Training Report

Data Overview:

- Dataset: 192,368 rows
- Text length varies between 200 to 3000 characters
- Multi-label tags column, up to 50 most frequent labels retained

Cleaning Steps:

- Removed punctuation, lowercased text
- Tokenization using NLTK
- Removed stopwords

Feature Engineering:

- Used TF-IDF Vectorizer
- Max features: 20,000
- Applied on cleaned title + text

Multi-label Transformation:

- Used MultiLabelBinarizer on tags
- Top 50 tags selected based on frequency

Model Used:

- LogisticRegression(solver='liblinear')
- Wrapped with OneVsRestClassifier

Training Performance:

Metric	Score
Hamming Loss	0.081
Precision	0.81

Recall	0.78
F1-score	0.79

Tools Used:

- Colab (for training and EDA)
- Gradio (for UI)

Visualizations:

- Bar plot of tag frequencies
- WordCloud of common terms

Conclusion:

- The model performs well for top-50 tag classification
- Real-time predictions are fast and interpretable
- Can be extended to larger datasets or news aggregators