Classi News – Intelligent News Categorization System

ADVANCED MACHINE LEARNING & NLP PROJECT

INTRODUCTION

- This project classifies news articles into World, Sports, Business, Science/Tech categories.
- Uses machine learning and Natural Language Processing (NLP) techniques.
- Dataset used: AG News from Kaggle.
- Built using Python, TF-IDF, and Multinomial Naïve Bayes.
- Includes visualizations and an interactive Gradio Web App.

Problem Statement

- • Huge volume of news articles daily
- • Manual classification is slow & inconsistent
- • Need: Automated, accurate, real-time categorization
- • Improves readability, personalization & media workflow

OBJECTIVES

- Clean and preprocess text using NLP.
- Train a robust TF-IDF + Naive Bayes model.
- Provide users an interactive interface to classify articles.
- Visualize dataset statistics and model performance.
- Store prediction history and export results.

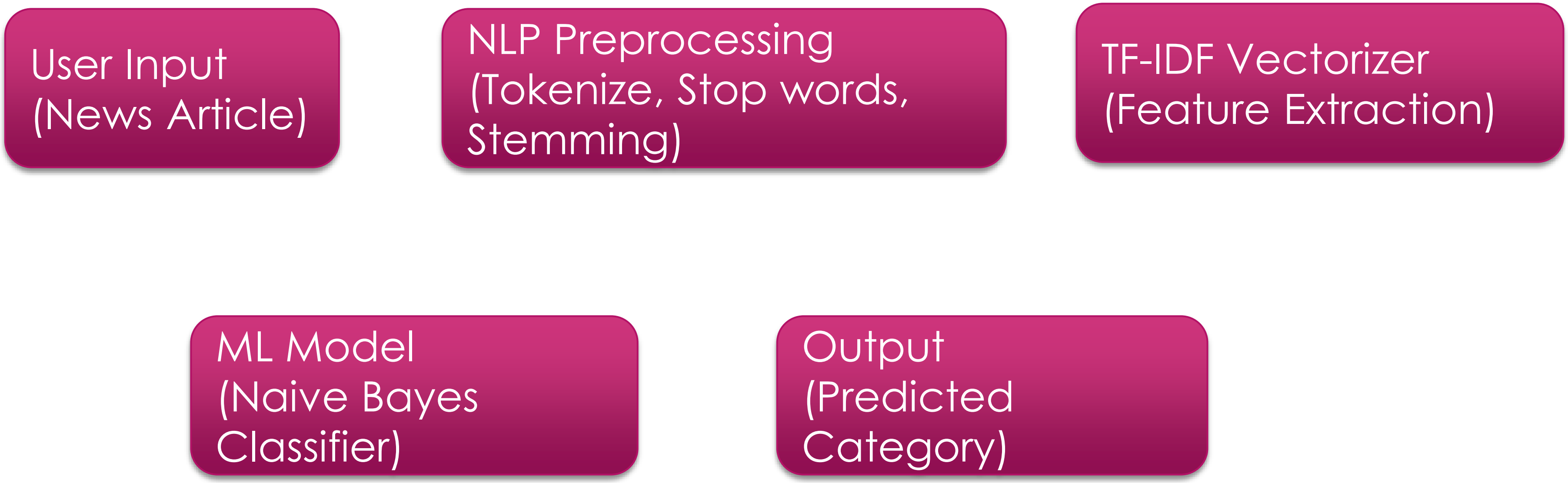
DATASET DETAILS

- Contains 120,000+ training news articles.
- Four predefined categories: World, Sports, Business, Sci/Tech.
- Each record includes: Title + Description.
- Dataset is well-balanced across categories.
- Used separate train and test CSV files.

Data Preprocessing

- Convert text to lowercase.
- Remove special symbols, numbers & punctuation.
- Remove stopwords using NLTK.
- Combine Title + Description → Text field.
- Create Processed_Text, Word_Count, Text_Length features.

System Architecture (Diagram)



User Input (News Article)

NLP Preprocessing (Tokenize, Stop words, Stemming)

TF-IDF Vectorizer (Feature Extraction)

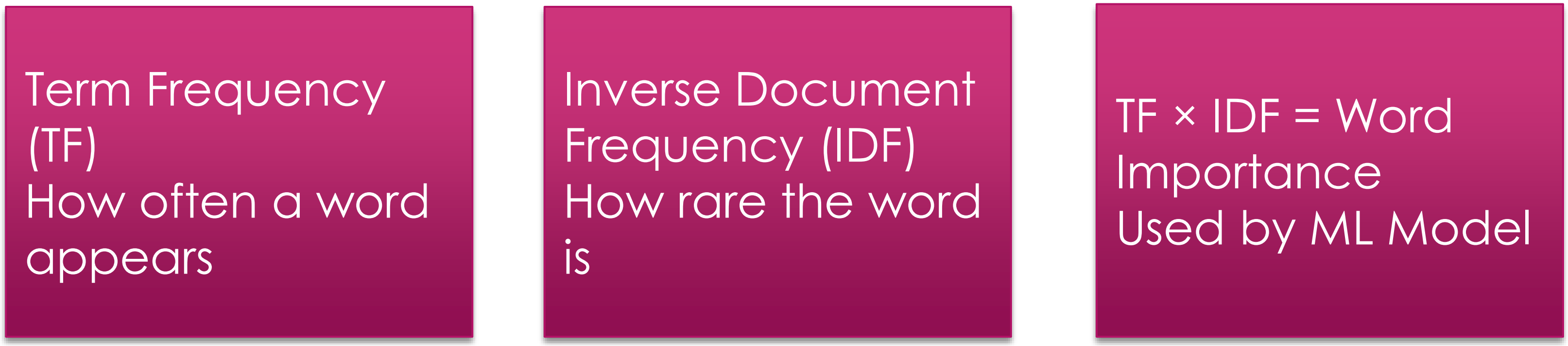
ML Model (Naive Bayes Classifier)

Output (Predicted Category)

TFIDF VECTORIZATION

- Convert text into numerical representation.
- Used 10,000 features with unigrams + bigrams.
- Removed rare/overly frequent words using min_df & max_df.
- Output: Sparse matrix for ML model.
- Improves model accuracy and generalization.

TF-IDF Working Diagram



Term Frequency (TF)
How often a word appears

Inverse Document Frequency (IDF)
How rare the word is

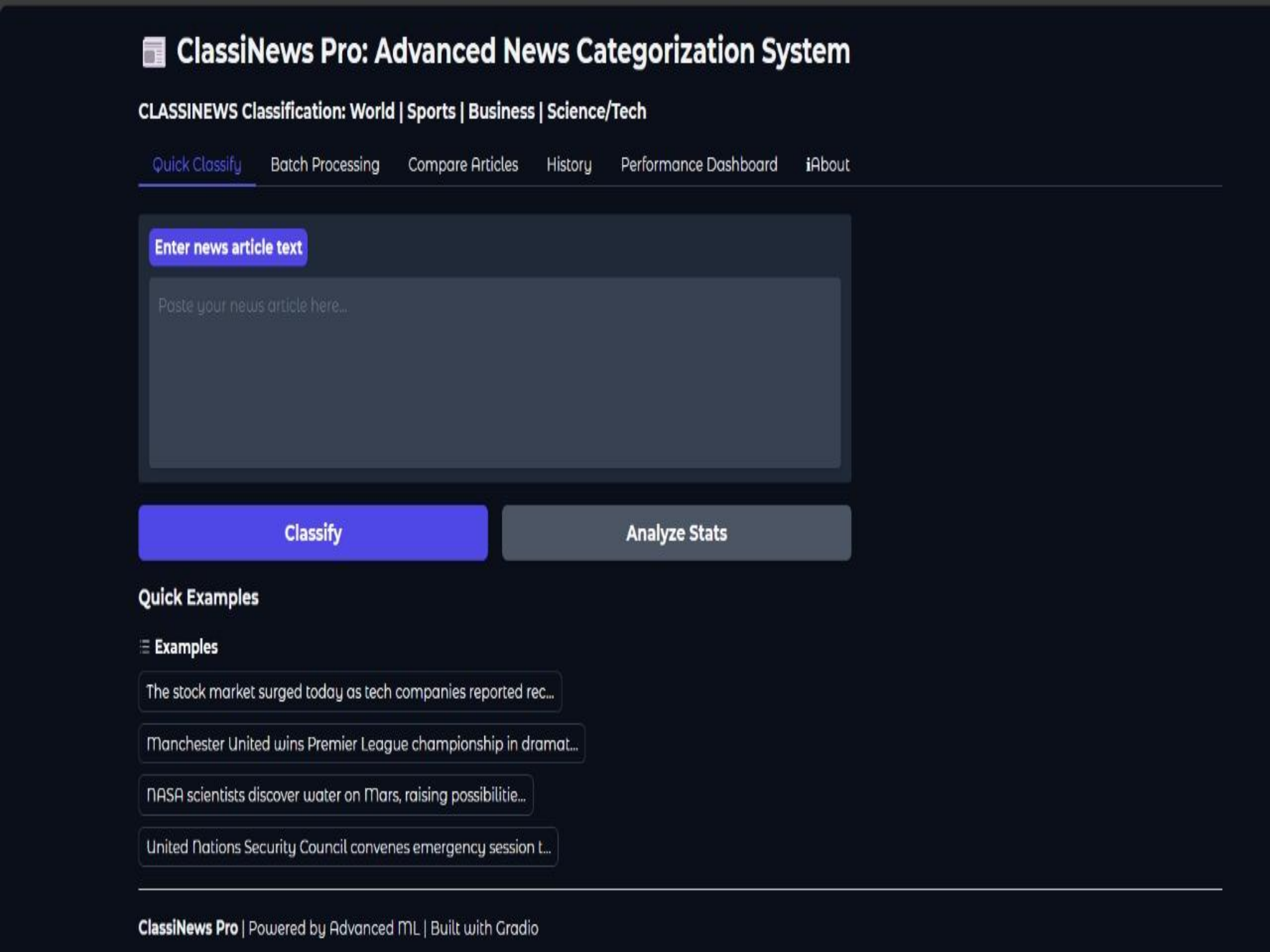
TF × IDF = Word Importance
Used by ML Model

Model Training

- • Train-test split (80/20)
- • Model: Multinomial Naive Bayes (best for text)
- • Learns patterns from TF-IDF vectors
- • Evaluated using accuracy, precision, recall, F1-score

INTERACTIVE WEB INTERFACE

- Interactive Web ApplicationUser can paste any news article to classify.
- Shows category + confidence bars.
- Tabs included:
- Quick Classify
- Batch Processing
- Compare Articles
- History Export
- Simple, clean UI for real-time predictions.



📰 ClassiNews Pro: Advanced News Categorization System

CLASSINEWS Classification: World | Sports | Business | Science/Tech

Quick Examples

Examples

The stock market surged today as tech companies reported rec...

Manchester United wins Premier League championship in dramat...

NASA scientists discover water on Mars, raising possibilitie...

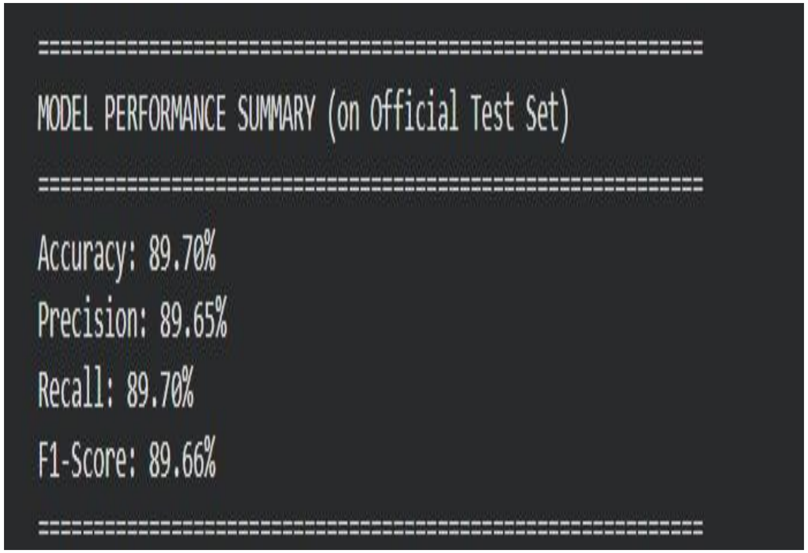
United Nations Security Council convenes emergency session t...

ClassiNews Pro | Powered by Advanced ML | Built with Gradio

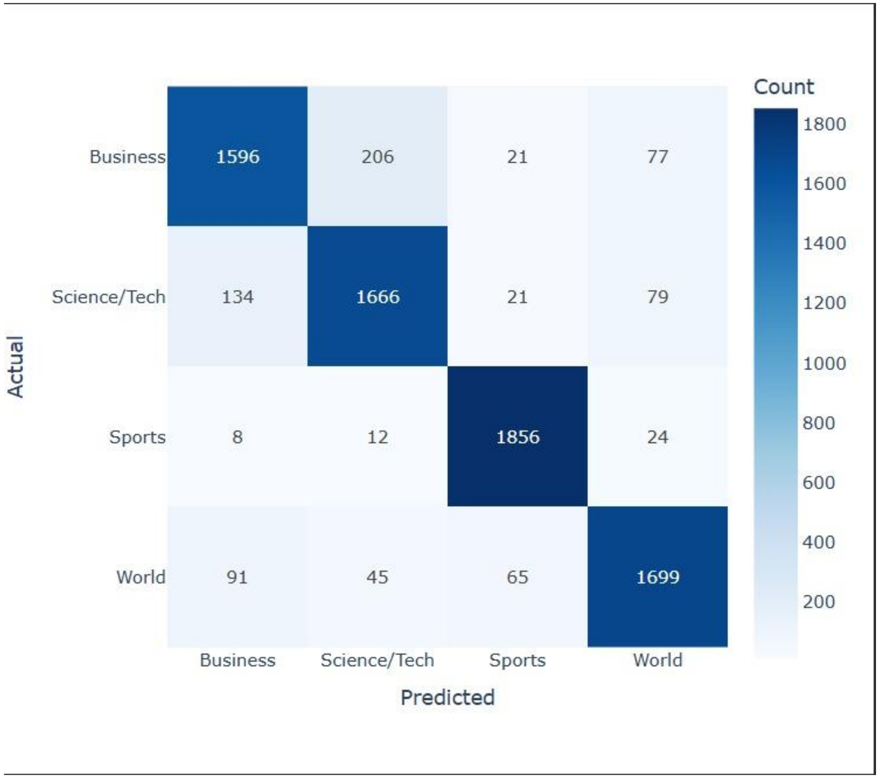
Results & Performance

- Accurate news classification system.
- User-friendly interactive interface.
- Strong model performance with TF-IDF + Naive Bayes.
- Good visualization and interpretability.
- Useful for real-world automated news categorization.

ACCURACY AND CONFUSION MATRIX



```
============================================================
MODEL PERFORMANCE SUMMARY (on Official Test Set)
============================================================
Accuracy: 89.70%
Precision: 89.65%
Recall: 89.70%
F1-Score: 89.66%
============================================================
```



Actual \ Predicted	Business	Science/Tech	Sports	World
Business	1596	206	21	77
Science/Tech	134	1666	21	79
Sports	8	12	1856	24
World	91	45	65	1699

Future Enhancements

- • Upgrade to BERT / Transformer models
- • Multilingual classification (Telugu, Hindi, Tamil…)
- • Personalized recommendations
- • Real-time news feed integration
- • Continuous learning through user feedback

Conclusion

- ClassiNews automates the entire news categorization workflow using
- NLP + TF-IDF + Machine Learning. It delivers accuracy, speed, and
- improved user experience for modern digital media platforms.