NewsBot Intelligence System 2.0 – Technical Documentation

Author: Mohanad Yassin– SoloTeam – ITAI2373

1. **Introduction**

NewsBot 2.0 is an advanced NLP-powered system designed to analyze, classify, summarize, and interpret news articles.

This documentation explains the system architecture, modules, NLP techniques, and implementation details.

2. **System Overview**

The system processes raw news articles and produces:

- Category classification

- Topic modeling insights

- Summaries

- Semantic search results

- Named entity recognition

- Multilingual translation support

3. Pipeline Architecture

1. Input – Raw articles

2. Preprocessing – Cleaning, stopword removal, lemmatization

3. Translation – Converts non-English text to English

4. NLP Modules:

- TF-IDF + Logistic Regression (classification)

- LDA Topic Modeling

- Transformer Summarizer

- Semantic Search with embeddings

- Named Entity Recognition

5. Output – Combined insights

4. Dataset & Model

Uses:

- newsbot_dataset.csv

- newsbot_model.joblib

- tfidf_vectorizer.joblib

5. Module Descriptions

Preprocessing: Cleans text for better analysis.

Classification: Predicts category of article.

Topic Modeling: Detects main themes.

Summarization: Produces short, readable summaries.

Semantic Search: Searches by meaning, not keywords.

Multilingual Layer: Detects language and translates.

Conversational Interface: Allows natural language queries.

6. Implementation Summary

Built with:

- Scikit-learn

- SpaCy

- HuggingFace Transformers

- Sentence Transformers

- Langdetect

- Google Translate API

7. Conclusion

NewsBot 2.0 is a full NLP pipeline that transforms raw text into actionable insights.