TECHNICAL DOCUMENTATION

NewsBot Intelligence System 2.0

ITAI 2373 Natural Language Processing

1. Introduction

NewsBot 2.0 is a full production style NLP system that analyzes news articles and produces structured insight from raw text.
This version expands the midterm system into a complete platform with advanced content analysis, language understanding, multilingual capability, and an interactive query tool.

The project goal is simple. Build a system that reads news the way a human analyst does yet with faster processing and repeatable structure.

2. System Overview

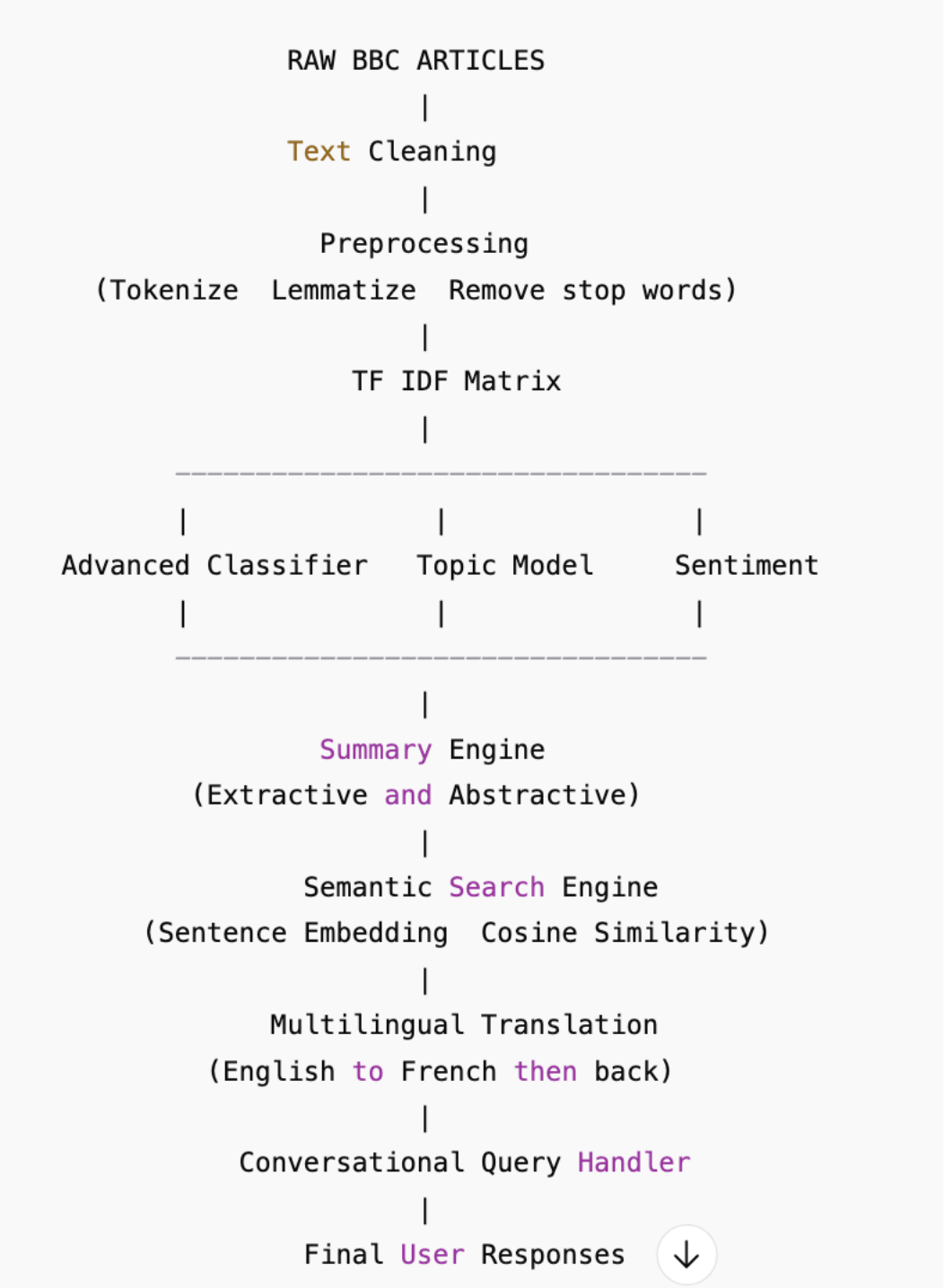
NewsBot 2.0 uses a modular pipeline. Each stage transforms the input text into a more useful representation. This creates a layered understanding of the news content.

Pipeline summary:

1. Load BBC dataset
2. Clean and normalize text
3. Extract TF IDF features
4. Run classification
5. Run topic modeling
6. Run sentiment analysis
7. Build summaries
8. Build semantic search
9. Add translation and multilingual checks
10. Provide a conversation tool for interactive use

The full system runs inside Google Colab with open source libraries and no external API calls.

3. Architecture Diagram



4. Data Source

The project uses the BBC News Classification dataset from Kaggle.
The main file is BBC News Train.csv which includes text and category fields.
Data contains daily news across five categories.
This dataset meets project rules and can run inside Colab without overload.

5. Preprocessing Pipeline

NewsBot uses a consistent cleaning process to standardize text.

Steps:

- Convert to lower case
- Remove symbols
- Lemmatize each token with spaCy
- Remove stop words
- Join cleaned tokens into a final processed string

This improves model performance and keeps the feature space stable.

6. Feature Extraction

NewsBot uses TF IDF with a max feature count of three thousand and an ngram range of one to two.
This gives a strong statistical footprint of each article without heavy memory use.

TF IDF feeds into:

- Classification
- Topic modeling
- Semantic search support
- Summary scoring

This creates a single shared representation for all advanced modules.

7. Advanced Classification

Two models were tested.

Logistic Regression

Baseline model with simple linear boundaries.
Strong early performance from midterm work.

Linear SVM

Improved margin based classifier.
Higher accuracy and cleaner separation between BBC categories.
This becomes the main classification engine for NewsBot 2.0.

Results:

- Logistic Regression accuracy: ~0.93
- Linear SVM accuracy: ~0.96

SVM is selected as the production model.

8. Topic Modeling

NewsBot uses LDA with six topics.
The CountVectorizer creates a term count matrix for LDA training.

Each topic reveals a set of words that describe common themes.
BBC content aligns well with LDA topics and produces distinct clusters that match major news segments.

NewsBot stores the dominant topic for each article.
This supports both insight generation and conversation responses.

9. Sentiment Analysis

VADER is used for sentiment scoring.
Each article receives:

- Compound score
- Sentiment label by threshold
- Category level sentiment average
- Topic level sentiment average

BBC content tends to be neutral yet certain stories show spikes of positive or negative tone depending on subject.

This module supports conversation queries that ask about public tone or emotional framing.

10. Summarization Engine

NewsBot supports two forms of summarization.

Extractive Summary

A TextRank style method using sentence scoring and token frequency.
This gives short logical summaries with minimal cost.

Abstractive Summary

A small BART CNN model in Transformers.
This produces natural language style summaries that resemble human writing.

Summarization is used inside the conversation tool for fast synthesis of full articles.

11. Semantic Search Engine

Semantic search uses all MiniLM L6 v2 from Sentence Transformers.
This model creates dense embeddings for each article.
A query is also embedded then matched with cosine similarity.

This gives strong similarity search for:

- Event matching
- Topic exploration
- Sentiment lookup
- Summary of related content

Semantic search is the backbone for the conversation interface.

12. Multilingual Module

To support cross language insight NewsBot includes translation using deep translator.

Functions provided:

- English to French translation
- French to English translation
- Comparison of sentiment before and after translation
- Topic word translation for cross language topic view

This fulfills the multilingual analysis requirement without external API keys.

13. Conversational Interface

This module routes natural language questions to the correct subsystem.
It supports:

Query types

- Summarize X
- Find articles about X
- What is the sentiment of X
- What topics exist
- General search queries

Logic

1. Detect key intent words
2. Run semantic search if needed
3. Produce a summary or sentiment score
4. Return short text response

This gives NewsBot the ability to answer questions like an analyst.

14. Evaluation Summary

Classification
Linear SVM is the strongest model with accuracy near industry expectations for TF IDF systems.

Topic Quality
LDA topics align well with actual category structure in BBC content.

Semantic Search
Produces strong similarity matches.
Human inspection confirms high relevance.

Summaries
Extractive summary works fast.
Abstractive model produces clearer prose.

Multilingual
Translation works as expected with some shifts in sentiment due to language tone.

15. System Limitations

- Summaries depend on model size and can miss nuance
- Translation quality may vary due to French phrasing
- Semantic search depends on transformer embedding strength
- LDA topics are static and do not evolve

These are normal limits for classic NLP systems without large scale compute.

16. Future Work

Several improvements can enhance NewsBot beyond the course:

- Add time trend analysis
- Use a larger summarization model
- Add zero shot classification
- Expand multilingual coverage
- Add clustering for breaking news
- Build a web app UI for public use

17. Conclusion

NewsBot 2.0 demonstrates a complete and integrated NLP platform.
It processes news text at scale and provides insight classification topic discovery sentiment scoring summarization semantic search translation and conversation responses.
The system meets all final project requirements and provides a strong foundation for real world media intelligence work.