Final Project

The NewsBot Intelligence System Capstone Project represents the full evolution of what started as a high accuracy midterm prototype into a real, production ready intelligence platform. The earlier version already hit an impressive milestone with a strong 97.99 percent classification accuracy using traditional Machine Learning methods. The goal for the final project was not just to maintain this performance, but to upgrade the entire system with real Software Engineering structure and add advanced Unsupervised Learning features that make it useful outside the classroom. The final platform is built to automatically analyze global news data, extract meaningful patterns, and deliver strategic insights that would help with proactive market monitoring or competitive intelligence.

A major part of this project involved cleaning up and rebuilding the original codebase so it could function like an actual application rather than a one off notebook. The focus was on organization, speed, scalability, and overall professional quality.The original midterm code was monolithic and tied directly to a single Jupyter Notebook. For the capstone, everything was broken out into a modular Python project. All repeatable logic such as preprocessing, feature engineering, and model training was moved into separate, well documented scripts like preprocessing.py and models.py. This follows the Single Responsibility Principle and makes the system far easier to read, maintain, and test.

The project was also restructured into a clean directory layout with a proper requirements.txt file so the environment can be fully recreated without extra setup. Several upgrades were made to improve efficiency and make the system deployable:

Model Persistence:
The trained classification pipeline, including the TF IDF Vectorizer and the Logistic Regression model, was serialized using joblib. This allows the system to load instantly without reprocessing the entire dataset or retraining the model.

API Integration:
All core analysis functions such as classification, sentiment analysis, and named entity recognition were wrapped inside a FastAPI service. This exposes the intelligence of the system through a fast REST API that can be connected to web apps, dashboards, or other tools like Streamlit or Gradio.

Efficiency Optimizations:
Batch processing with spaCy's nlp.pipe method was used to reduce runtime during POS tagging and NER, which is especially helpful when dealing with large streams of incoming news articles.

The biggest functional addition in the final system is the Topic Modeling module. Unlike the supervised classifier, this module uses Unsupervised Learning to uncover hidden patterns and themes in the news corpus. While the supervised model predicts one of the five categories from Modules 3, 4, and 7, the Topic Modeling module looks for underlying themes that appear naturally across documents. For example, even if the classifier labels a story as Technology, the LDA model might reveal a deeper theme like global chip shortages, based on repeated terms such as chip,

supply, fab, Taiwan, and demand. This gives a deeper level of insight that goes beyond simple labeling.

The Topic Modeling workflow is built with the Gensim LDA library and produces several outputs that translate directly into Business Intelligence:

Data Preparation:
The processed corpus is converted into a numerical bag of words representation along with a dictionary. This format is required for LDA and ensures fast, efficient computation.

Topic Interpretation:
For each discovered theme, the system extracts the top ten most influential words. These word clusters make it easier for analysts to understand what each theme represents. For example, a set of words like salary, contract, transfer, team, and agent may point to the theme of professional sports finance.

Interactive Visualization:
Using pyLDAvis, the model generates an interactive visualization where themes appear as circles that users can explore. This tool makes complex patterns clear and is extremely helpful for people who are not familiar with NLP.

With this extension, NewsBot does more than classify articles. It now reveals emerging themes, giving users early warning on trends that might not be obvious through simple labels.

The NewsBot Intelligence System Capstone Project represents a full life cycle build, showing both strong data science skills and solid software engineering principles. By upgrading the midterm system, improving its efficiency, and adding Topic Modeling for deeper strategic insight, the final product is now a deployable, reliable, and versatile intelligence platform.

The system can quickly process unstructured text, detect patterns, and turn raw news into actionable insights. It gives organizations the ability to automate media monitoring, spot early market shifts, and gain a competitive edge through data driven intelligence.