NewsBot Intelligence System 2.0 — Technical Report

Yunze Wu | ITAI 2373 – Natural Language Processing | Houston Community College

The NewsBot Intelligence System 2.0 extends the midterm NewsBot pipeline into a more advanced and production-oriented system capable of classification, topic modeling, sentiment analysis, summarization, semantic search, multilingual processing, and interactive querying. The system uses the BBC News dataset, which contains articles across five categories: business, politics, sports, entertainment, and technology. The goal of this project is to demonstrate mastery of end-to-end NLP workflows and produce a system that resembles real-world media intelligence applications.

The system begins with preprocessing, which standardizes the text through lowercasing, punctuation removal, stopword filtering, and lemmatization. This step improves the quality of all downstream modules. TF-IDF vectorization is used to transform the cleaned corpus into a high-dimensional feature space suitable for machine learning models and semantic retrieval. Several classifiers were trained, including Logistic Regression, Linear SVC, and Random Forest. Logistic Regression provided the strongest baseline performance on sparse TF-IDF features. A small hyperparameter tuning process was applied to Logistic Regression using grid search, leading to further improvement and greater model stability.

A topic modeling module was implemented using Latent Dirichlet Allocation (LDA). This allowed the system to reveal latent semantic patterns within the news corpus. By examining the top contributing words for each topic and comparing topic distributions across categories, the model demonstrates how news themes naturally cluster, such as financial vocabulary appearing in business articles or competitive/action-focused terms appearing in sports content.

Sentiment analysis was performed using VADER to compute compound sentiment scores for each article. These scores were converted into categorical sentiment labels and aggregated across news categories. The resulting distribution provides insights into how tone varies across domains—politics tends to be more neutral, entertainment more positive, and certain business or tech reports fluctuate depending on the nature of the content. These visualizations would be directly useful for a media analytics dashboard.

Two major language understanding and generation features were added. First, extractive summarization, implemented through a frequency-based scoring of sentences, enables the system to produce concise overviews of longer articles. Second, a semantic search function uses cosine similarity over TF-IDF vectors to retrieve articles most relevant to a user query. This turns the system into a mini search engine capable of retrieving semantically related content rather than relying on keyword overlap alone.

To demonstrate multilingual adaptability, the system integrates a translation component. Articles can be translated into another language (Chinese was used in this project) and translated back into English. The sentiment and content of the back-translated text are reanalyzed, allowing examination of how translation affects meaning. This showcases how the pipeline could be extended to handle international news sources by combining translation with existing NLP modules.

A simple conversational interface was implemented to provide interactive access to the entire system. Users can enter free-form commands such as “search: <query>” or “summary: <index>” to retrieve articles, generate summaries, inspect sentiment, or display article content. This creates a foundation for future deployment as a web application or chatbot.

Model evaluation demonstrates that Logistic Regression remains the strongest classifier after tuning. LDA topics align well with expectations for BBC reporting. Semantic search retrieves articles accurately and consistently. Summarization is coherent for most samples. Sentiment analysis performs reasonably but remains limited by VADER's domain-general nature. Translation introduces some semantic drift, which is useful to observe in multilingual evaluation.

Throughout development, several challenges emerged. Sparse TF-IDF features require careful model selection; LDA topic coherence is sensitive to preprocessing choices; translation APIs produce inconsistent results; and interactive input is limited in the Colab environment. System limitations include extractive summarization's inability to rewrite or condense ideas, sentiment analysis that is not domain-adapted, and the absence of neural embedding models that would improve search and classification.

Future enhancements include adopting transformer-based summarization (e.g., BART or T5), replacing TF-IDF with dense embeddings such as Sentence-BERT, adding multilingual classification through mBERT or XLM-R, deploying the system using FastAPI with a user-friendly interface, and integrating personalized recommendation components.

In conclusion, the NewsBot Intelligence System 2.0 presents a cohesive and extensible news analysis pipeline. It integrates classification, topic modeling, summarization, sentiment analysis, semantic retrieval, multilingual processing, and interactive dialogue capabilities. This project demonstrates a comprehensive understanding of modern NLP techniques and how they can be combined to create meaningful, real-world applications in media monitoring and news intelligence.

Appendix (Code References)

https://github.com/YuznW/portfolio