Jeffery Dirden
ITAI 2373
August 4, 2025

Technical Doc

Project Overview

The NewsBot Intelligence System is an AI-powered news aggregation and analysis tool designed to retrieve real-time news articles from the web, extract meaningful information such as named entities and sentiment, classify articles by category, and present comprehensive insights through an interactive interface. This system supports informed decision-making and efficient media monitoring by combining web search, natural language processing, and machine learning classification.

System Architecture

The system architecture consists of several modular pipeline components. First, the News Retrieval module uses the DuckDuckGo Search API to fetch the latest news articles based on user queries. Next, the Text Preprocessing component cleans and prepares raw text data to optimize analysis accuracy. The Entity Extraction module employs spaCy's NLP model to identify named entities such as persons, organizations, and locations within news content. Sentiment Analysis is conducted using NLTK's VADER sentiment lexicon to determine the polarity of the articles. The Classification and Insights module utilizes a trained Logistic Regression model on TF-IDF vectorized data to categorize articles into

predefined classes. Finally, a User Interface built with Gradio provides an accessible way for users to query and explore enriched news content.

Core Components

The News Retrieval module is implemented via a function that leverages the DuckDuckGo Search API, allowing the system to collect recent news snippets according to specified search queries. It returns metadata including article titles, summaries, and sources.

The Natural Language Processing module relies on spaCy's English language model, which provides tokenization, named entity recognition, and other linguistic analyses.

Entity Extraction takes concatenated title and summary text as input and outputs detected named entities with labels such as PERSON, ORG, and GPE, enriching the article metadata.

Sentiment Analysis uses NLTK's SentimentIntensityAnalyzer to generate sentiment polarity scores, including positive, negative, neutral, and a compound score reflecting overall sentiment strength.

For classification, feature engineering is performed using TF-IDF vectorization on preprocessed text data. A Logistic Regression model trained on a curated dataset predicts categories such as Politics, Sports, Technology, Business, and Entertainment. Dummy feature vectors were initially employed for compatibility with the model, with provisions for future feature expansion. The module also generates actionable insights by combining classification confidence, sentiment analysis results, and named entity recognition.

The User Interface is built using Gradio, providing a web-based platform for users to enter search queries and receive tabular outputs displaying article titles, summaries, sources, and recognized entities along with system status.

Development Process

The development followed a modular design approach to promote extensibility and maintainability. The integration of live web search with natural language processing components enabled dynamic data processing. Data cleaning, entity extraction, and sentiment analysis were implemented as reusable functions to maintain clarity and reusability. The classification model was trained and validated on a small but representative dataset. Extensive testing scripts were developed to cover diverse news categories and handle ambiguous cases. Finally, a user-facing Gradio interface was created to demonstrate the system's full capabilities interactively.

Testing and Evaluation

Testing involved a variety of news topics, including technology, sports, business, health, environment, and ambiguous multi-topic articles. The accuracy of entity extraction was validated using spaCy's outputs. Sentiment analysis results were cross-checked for alignment with the articles' tone. Classification confidence scores were reviewed to handle cases with moderate or low confidence, including user alerts and fallback mechanisms. Error handling was implemented to manage missing fields or incomplete search results, ensuring robust system performance.

Challenges and Solutions

One key challenge was handling ambiguous news articles that blended multiple topics. To address this, confidence thresholds were implemented to generate user alerts when predictions had moderate or low confidence. Another challenge was the limited labeled dataset for classifier training. To mitigate this, dummy feature vectors were used initially, with plans for dataset expansion to improve

accuracy. Additionally, variability in real-time web search results required the inclusion of error handling for empty or incomplete data to maintain stability.

Future Work and Improvements

Future improvements include expanding the training dataset to cover a wider range of categories and ambiguous cases. The dummy features will be replaced with real-time sentiment and metadata features to enhance classification accuracy. Integration of transformer-based language models, such as BERT, will be explored to improve contextual understanding. Temporal analytics will be added to track news trends over time. Incorporation of external knowledge bases like Wikipedia is planned to enrich entity descriptions. Lastly, user feedback loops will be introduced to enable active learning and continuous model refinement.

References

DuckDuckGo Search API via the duckduckgo_search Python package was utilized for news retrieval. The spaCy library and its English language model (en_core_web_sm) provided the foundation for natural language processing tasks. NLTK's VADER SentimentIntensityAnalyzer enabled sentiment scoring. The classification was performed using Scikit-learn's Logistic Regression model with TF-IDF features. Gradio was used for building the interactive user interface.