Technical Documentation

Project Title: NewsBot Intelligence System 2.0

Student: Rubén Darío Valenzuela

Course: ITAI 2373 – Natural Language Processing

System Overview

The NewsBot Intelligence System 2.0 is an advanced natural language processing (NLP) platform that classifies, summarizes, and analyzes news content using state-of-the-art AI techniques. It builds on the midterm version by incorporating multilingual support, advanced classification, sentiment analysis, summarization, and conversational interface features.

System Architecture

The system pipeline includes the following modules: data ingestion, preprocessing, feature extraction, classification, sentiment/emotion detection, summarization, and translation. Each module operates independently, making the pipeline modular and scalable.

Technologies Used

• Python • scikit-learn • pandas • NLTK • spaCy • VADER • TextBlob • Hugging Face Transformers • Google Translate API • Jupyter Notebook • GitHub

Model Performance

Using the BBC News dataset, the system achieved 97% classification accuracy across five categories: politics, tech, business, sports, and entertainment. The TF-IDF vectorizer and Multinomial Naive Bayes classifier formed the core of the categorization engine.

Challenges & Solutions

Key challenges included ensuring accurate multilingual translations, maintaining high classification accuracy, and integrating multiple NLP modules. These were addressed using robust libraries and iterative testing. Troubleshooting model drift and balancing precision and recall were essential steps.

Scalability & Deployment

The modular structure allows easy deployment as a web application or REST API. With a command-line interface and potential frontend integration, the system is adaptable for journalism, social monitoring, and educational use.

Advanced Extensions

To explore beyond the course requirements, the NewsBot Intelligence System 2.0 integrates the multilingual capabilities of mT5 and explores the application of RoBERTa for more refined sentiment classification. Additionally, we designed the pipeline to be adaptable for fake news detection and electoral media analysis, offering real-world relevance and ethical insight into AI deployment. These extensions demonstrate a research-oriented mindset and lay the groundwork for future improvements such as multimodal analysis or domain-specific fine-tuning.

Conclusion

This project showcases a production-ready NLP system that integrates multiple AI tools to analyze, classify, and summarize multilingual news content. It reflects a high level of technical understanding and practical application of AI methods, contributing to the advancement of intelligent information systems.