Report : NewsBot Intelligence System: A Comprehensive NLP Pipeline for News Classification

Our team, Group 5, comprised Franck, Iman, and Kimberly, collaborated closely to develop and implement the NewsBot Intelligence System. Throughout the project, we held regular brainstorming and code review sessions to align our contributions and resolve blockers. Franck focused heavily on model integration and data engineering, Iman managed visualization, and Kimberly oversaw documentation and presentation. We used Google Colab for coding collaboration and GitHub for version control. Communication was facilitated via group messaging, and task allocation was divided based on individual strengths, ensuring efficiency and mutual learning. The success of our project heavily relied on each team member's active engagement and willingness to support one another through complex stages of the pipeline.

We processed 1,490 news articles spanning five categories: Business (336), Tech (261), Politics (274), Sports (346), and Entertainment (273). Our analysis included TF-IDF vectorization, part-of-speech (POS) tagging, syntactic parsing, and sentiment analysis. We visualized linguistic differences using category-specific POS heatmaps and TF-IDF bar charts. For example, proper nouns (NNP) were most frequent in Sports and Tech, indicating high mention of specific entities and innovations. Entertainment emerged as the most positive category (mean sentiment score: 0.7043), while Politics was the most negative (mean sentiment score: 0.1250). These patterns helped identify emotional tone and language usage that could inform editorial strategy.

Integrating multiple NLP tasks into one unified pipeline posed several challenges. First, ensuring compatibility across tools like SpaCy, Scikit-learn, and NLTK required careful dependency management. Second, extracting consistent features (e.g., TF-IDF, sentiment, syntactic structure) from heterogeneous content required significant preprocessing. Franck addressed these by standardizing input formats and debugging transformation pipelines. Third, our initial model training failed due to stratified sampling errors on small datasets. We resolved this by scaling to 1,490 articles and ensuring balanced category representation. Finally, computational limitations in Colab led us to optimize code for memory and runtime, especially during TF-IDF matrix handling and classification stages.

The NewsBot Intelligence System addresses real business needs in media, marketing, and analytics. Its automated article classification and sentiment tagging can support editorial prioritization, content moderation, and reader engagement tracking. For example, media outlets could use sentiment insights to tailor content recommendations or monitor political bias. The rich entity recognition system allows journalists or researchers to quickly identify people, organizations, and dates mentioned in articles, facilitating fact-checking and relationship mapping. Furthermore, the classification engine—despite an initial accuracy of 40%—has potential to scale with additional training data and feature engineering, paving the way for commercial deployment in content management platforms.

- Franck developed and tested the classification model, designed the full preprocessing pipeline, and integrated sentiment analysis tools.
- Iman focused on visualization (heatmaps, distribution plots), final presentation formatting, and led the NER implementation.
- Kimberly contributed to research, coordinated document editing, and managed project deadlines and compliance with academic criteria.

Our balanced distribution of roles enhanced both efficiency and quality, and each member cross-reviewed one another's work.

To improve the system, we plan to:

- Fine-tune classification with larger and more diverse datasets.
- Apply deep learning (e.g., BERT) for improved sentiment/context understanding.
- Enable real-time data scraping and analysis.

This project deepened our understanding of natural language processing and full-stack AI system development. We applied theoretical knowledge in a practical, business-oriented context. Franck honed skills in machine learning workflows and data engineering. Iman gained expertise in linguistic annotation and data visualization. Kimberly developed technical writing, research, and team coordination abilities. The experience also improved our proficiency in collaborative tools: GitHub, Google Colab. Overall, this project strengthened our readiness for real-world roles in AI, data science, and software development. Our GitHub link is below.

https://github.com/imid12/miniature-eureka-Group5/tree/main/ITAI2373-NewsBot-Midterm