

Executive Summary

Group 5 ITAI 2373

Franck Kolontchang - Iman Haamid - Kimberly Navarrete - Marvin Azuogu

Github Link:

https://github.com/imid12/miniature-eureka-Group5/tree/main/ITAI2373-NewsBot_2.0-Final

- The NewsBot Intelligence System 2.0 is a comprehensive NLP system detailed in a Jupyter Notebook. It handles text preprocessing, Named Entity Recognition (NER), sentiment and emotional tone analysis, article classification, and feature extraction using TF-IDF. The system aims to provide business insights for media monitoring and market research.
- Setup requires several Python libraries like `spacy`, `scikit-learn`, and `nltk`, which are installed along with essential linguistic models and data.
- The notebook loads news datasets, typically from CSV, and explores them by displaying statistics, checking for missing values and duplicates, and analyzing text length distribution before and after preprocessing.
- A robust preprocessing pipeline cleans text by removing HTML tags, URLs, and special characters, then tokenizes, removes stop words, and lemmatizes it for consistency and efficiency.
- Feature extraction utilizes TF-IDF to convert text into numerical features, highlighting significant terms. This is then used to visualize top terms per category, generate word clouds, and create bar charts and heatmaps comparing TF-IDF scores across categories.
- Additionally, the document outlines an "Advanced Content Analysis Engine" with enhanced classification (multi-label and hierarchical), topic modeling (LDA, NMF, BERTopic), and nuanced sentiment analysis (intensity, specific emotions, aspect-based).
- The "Language Understanding & Generation" section covers text summarization (extractive and abstractive), semantic search (understanding meaning and intent), and content enhancement (cross-referencing, knowledge graph integration, fact

augmentation).

- The "Multilingual Intelligence" section details cross-language analysis (detection, multilingual embeddings, comparative insights) and translation integration (on-demand, pre-translation, quality assurance).
- Finally, the "Conversational Interface" enables natural language queries (NLU, dialogue management, query rewriting) and interactive exploration, allowing users to refine searches, drill down into insights, and receive actionable responses.