Jeffery Dirden
ITAI 2373
August 4, 2025

Reflective Journal

As students in the ITAI2373 course, this project has been a pivotal opportunity to bridge theoretical knowledge of Natural Language Processing (NLP) with real-world, hands-on experience. Our goal was to build a NewsBot Intelligence System capable of analyzing and extracting key information from a large corpus of news articles. Through collaborative effort, technical research, trial and error, and iterative improvements, we successfully implemented a comprehensive NLP pipeline including data preprocessing, tokenization, stopword removal, TF-IDF vectorization, POS tagging, syntactic parsing, and named entity recognition.

Jeffery, acting as the lead architect and analyst, was responsible for designing and implementing the core structure of the application. He handled the majority of the technical development including the data loading pipeline, the preprocessing stages using spaCy and NLTK, and the transformation of text using TF-IDF vectorization. He also conducted the syntactic analysis using part of speech tagging, generated word clouds and heatmaps, and led the deep dive into entity recognition and term frequency analysis across categories. In addition, Jeffery designed and integrated the Gradio-based frontend which allows users to interactively input news headlines or full content and receive real-time feedback, visualizations, and predictions. This frontend was a critical

component of the system's usability and made the model outputs accessible and interpretable for nontechnical users.

Williane contributed during the research and design phases. She assisted with selecting relevant libraries and understanding how to apply core NLP tools within the context of the dataset. She also reviewed and tested parts of the pipeline during implementation, especially in the early stages, to ensure clarity and alignment with our academic goals. While her technical contributions were limited, she played an important role in conceptual support and was consistently engaged in the learning process throughout development.

From a learning perspective, this project challenged both of us to apply a wide array of machine learning and natural language processing skills. We navigated the challenges of unstructured data and made strategic decisions on how to tokenize, clean, and vectorize text to enable downstream analysis. Working with TF-IDF taught us the value of weighting features based on document importance, and we saw firsthand how unigrams and bigrams could capture subtle but critical distinctions in language. The integration of part of speech tagging and named entity recognition allowed us to extract meaningful patterns and insights from complex linguistic structures. These components deepened our understanding of text structure and the role of syntactic and semantic features in data science.

Gradio served as the final layer in this project and gave us the opportunity to turn a technical pipeline into a user centered product. We learned how to deploy models in an interactive environment and saw the importance of user interface design when communicating data driven insights. It brought a sense of completeness to the system and made our work feel tangible and ready for real world application.

In conclusion, this project offered a rewarding experience that combined theory, application, collaboration, and creativity. We both grew significantly in our understanding of how NLP systems function in production and feel better prepared to contribute to similar projects in professional environments. The blend of technical depth, collaboration, and real time interactivity made this one of the most engaging projects of our academic journey so far.