MID-TERM GROUP PROJECT: NEWSBOT INTELLIGENCE SYSTEM

Course Information:

NLP ITAI 2373 - Midterm

Team Name:

FSM² (FSM Squared)

Team Members:

Faiza Abdullah

(https://github.com/AbdullahFaiza/NLP-ITAI2373/tree/main/ITAI2373-NewsBot-Midterm)

Sha'Rise Griggs

(https://github.com/stauriea21/itai2373-portfolio)

Professor:

Patricia Mcmanus

The NewsBot Intelligence System midterm project for ITAI 2373 was a comprehensive exercise in integrating NLP techniques from Modules 1-8 to build an end-to-end system for processing, categorizing, and analyzing news articles. This project required our team to combine text preprocessing, TF-IDF feature extraction, named entity recognition (NER), sentiment analysis, POS tagging, and topic modeling (via Latent Dirichlet Allocation, LDA) into a unified pipeline. The goal was to create a robust system for media monitoring, business intelligence, and content management, addressing real-world applications like those used by Google News or financial firms. Despite significant team collaboration challenges, the project deepened our understanding of NLP pipelines, system design, and ethical considerations, while highlighting the practical value of these techniques in industry settings.

Key Insights and Takeaways

This project was a transformative journey in applying NLP to solve real-world problems. The NewsBot system successfully integrated multiple NLP components, including a text preprocessing pipeline (Module 2), TF-IDF-based classification (Module 3, 7), NER (Module 8), sentiment analysis (Module 6), and topic modeling (Research Extension). The baseline classifier (RandomForest with TF-IDF) achieved solid performance, while the Research Extension's LDA implementation added semantic context, slightly improving accuracy for ambiguous articles (e.g., tech-business hybrids). The c_v coherence score (0.3–0.7 range) validated topic quality, and visualizations like topic distribution bar plots enhanced interpretability.

The project's real-world relevance became clear through its applications in media monitoring (e.g., categorizing news for Google News), market sentiment analysis (e.g., tracking company mentions for financial firms), and content management (e.g., organizing articles for media companies). The ability to extract entities (e.g., "Apple Inc.", "Tim Cook") and sentiments (positive/negative) provided actionable insights, such as identifying negative press for crisis management. The Research Extension’s topic modeling added depth, revealing thematic trends (e.g., economy-related topics in business articles), aligning with Module 1’s focus on NLP applications.

Technical Mastery: Most Challenging and Useful NLP Techniques

Most Challenging Techniques:

1. Topic Modeling with LDA (Research Extension): Implementing LDA using scikit-learn and gensim was challenging due to a ValueError: numpy.dtype size changed, may indicate binary incompatibility. This required debugging and reinstalling compatible versions (numpy==1.26.4, gensim==4.3.3), followed by a runtime restart in Colab. Tuning LDA parameters (e.g., number of topics) and interpreting the c_v coherence score (Module 7) were also complex, as small datasets (e.g., 5-article sample) led to lower coherence and less meaningful topics.
2. NER Integration (Module 8): Integrating SpaCy’s NER into the NewsBotIntelligenceSystem class required careful handling of entity types (e.g., PERSON, ORG) and ensuring the custom model (en_core_web_sm) worked with the pipeline. Ensuring robustness for informal or noisy text (e.g., social media-like articles) was difficult, echoing Lab 05’s Messy Text Challenge.

Most Useful Techniques:

1. TF-IDF Feature Extraction (Module 3): TF-IDF was the backbone of the classification pipeline, transforming raw text into numerical features for the RandomForest classifier. Its ability to weigh term importance (e.g., “economy” in business articles) drove high baseline accuracy, making it indispensable for text classification tasks.
2. Sentiment Analysis (Module 6): Using TextBlob for polarity scoring provided actionable insights (e.g., labeling “Apple’s new iPhone is innovative” as positive). Its

integration into the NewsBotIntelligenceSystem class enabled business-relevant outputs, such as flagging negative sentiment for PR teams.

These techniques, particularly TF-IDF and sentiment analysis, were critical for building a practical system, while LDA and NER added depth for advanced analysis, aligning with the project's goal of generating business value.

Technical Integration Challenges

Combining multiple NLP tasks into a cohesive pipeline was a significant challenge:

1. Dependency Management: The numpy/gensim incompatibility (ValueError) disrupted the LDA implementation. We resolved this by reinstalling compatible versions and restarting the Colab runtime, ensuring the CoherenceModel worked for topic evaluation. This mirrored Lab 07's DataFrame KeyError challenge, reinforcing the need for rigorous dependency checks.
2. Feature Integration: Combining LDA topic distributions with TF-IDF features required aligning sparse matrices (X_tfidf) with dense topic features (doc_topic_matrix). Using np.hstack ensured compatibility, but careful indexing was needed to avoid shape mismatches. This extended Lab 07's multimodal fusion challenge, where text and audio feature dimensions had to align.
3. Pipeline Modularity: Integrating NER, sentiment analysis, and topic modeling into the NewsBotIntelligenceSystem class required a modular design. I added optional parameters (e.g., topic_model, count_vectorizer) to support both baseline and enhanced modes, ensuring flexibility without breaking existing functionality.
4. Small Dataset Limitations: The fallback 5-article sample dataset (4 business, 1 tech) limited LDA's topic quality and classification improvements. I mitigated this by including a warning in the code and prioritizing the full BBC News Train.csv dataset when available, drawing on Module 2's preprocessing lessons for robust data handling.

These challenges honed debugging and system design skills, particularly in managing complex pipelines and ensuring compatibility across NLP components.

Business Value Assessment

The NewsBot Intelligence System addresses several real-world problems:

1. Media Monitoring: Automatically categorizes news articles (e.g., business, tech, politics), enabling platforms like Google News to organize content efficiently. The LDA extension enhances this by identifying thematic trends (e.g., "AI innovation" in tech articles).
2. Business Intelligence: Extracts entities (e.g., "Apple Inc.") and sentiments (e.g., negative for "Enron scandal"), helping financial firms track market sentiment or media companies monitor competitor coverage.
3. Content Management: Organizes large volumes of articles by category and topic, streamlining workflows for newsrooms or content aggregators.
4. Market Research: Analyzes public sentiment trends (e.g., positive sentiment for "new iPhone launch"), aiding product launches or policy analysis.

For example, a media company could use the system to flag negative articles about a brand (e.g., "WorldCom ex-boss" with negative sentiment) for crisis management, while a financial firm could track mentions of "Enron" for investment decisions. The LDA extension adds value by uncovering hidden themes (e.g., economic downturns), enhancing strategic insights.

Ethical Considerations

Automated news analysis poses several risks:

1. Bias Amplification: The classifier may reinforce biases in the training data (e.g., overrepresenting business articles in the sample dataset), leading to misclassifications

of minority categories like tech. We mitigated this by comparing baseline and enhanced models, but diverse datasets are needed for fairness.

2. Misinterpretation of Sentiment: Sentiment analysis may misclassify nuanced texts (e.g., sarcasm like "Great job, another outage!"), potentially misleading businesses. This echoes Lab 07's challenge with ambiguous social media posts.
3. Privacy Concerns: NER extracting names (e.g., "Bernie Ebbers") could raise privacy issues if applied to sensitive datasets. We ensured only public news data was used, but real-world applications must comply with regulations like GDPR.
4. Topic Misassignment: LDA may assign misleading topics to ambiguous articles (e.g., a tech-business hybrid), affecting downstream decisions. Qualitative analysis of sample articles helped identify such errors, but advanced models like BERTopic could improve accuracy.

These considerations highlight the need for transparency, diverse training data, and human oversight to ensure ethical deployment.

Future Learning / Enhancements

The project sparked interest in several NLP topics:

1. Transformer-Based Models: Exploring BERT or BERTopic for topic modeling and classification could capture contextual nuances better than LDA or TF-IDF.
2. Multi-Label Classification: Extending the system to handle articles with multiple categories (e.g., tech and business) would improve robustness for ambiguous texts.
3. Real-Time Analysis: Integrating social media data from platforms like X (Module 1) could enable real-time trend detection, building on social media monitoring application.
4. Bias Mitigation: Techniques like adversarial training could reduce cultural or dataset biases, addressing the ethical concerns.

We are excited to dive into transformer models and real-time NLP, as they align with industry trends in automated content analysis.

Team Collaboration Analysis

Challenges in Coordination: Our team (Faiza Abdullah, Sha'Rise Griggs, Kaden Glover, Marvin Azuogu) faced significant collaboration challenges. We started with setting pace for initial discussion, when no one participated, Faiza took lead and completed the notebook assignment and also did research extension bonus of LDA, Sha'Rise supplemnetd by doing other bonus segments like Research Extension (Emotional Intelligence) and Advance analysis (Temporal Sentiment). Kaden opted for report and despite asking him to attempt one bonus activity he only prepared baseline report. Marvin started very late and took up remaiing bonus areas but not deliver causing missing deadlines. This disrupted our timeline, and even though we wanted to do the video, we were unable to.

Individual Contributions:

Faiza Abdullah: Led the project, implementing the core NewsBotIntelligenceSystem class, integrating TF-IDF classification, NER, sentiment analysis, and LDA (Research Extension). Resolved technical issues (e.g., numpy/gensim incompatibility) and wrote comprehensive documentation. (https://github.com/AbdullahFaiza/NLP-ITAI2373/tree/main/ITAI2373-NewsBot-Midterm)

Sha'Rise Griggs: Contributed to bonus areas, including Emotional Intelligence (sentiment-driven insights), and Temporal Sentiment analysis. Assisted with qualitative evaluations and visualizations. (https://github.com/stauriea21/itai2373-portfolio)

Ensuring Quality: We used GitHub for version control with Faiza managing commits and Sha'Rise reviewing bonus sections.

Portfolio Value

We will present the NewsBot Intelligence System as a flagship project in our portfolio to demonstrate:

1. End-to-End NLP Pipeline: Showcase the integration of preprocessing, TF-IDF, NER, sentiment analysis, and LDA, highlighting my ability to build industry-relevant systems.
2. Problem-Solving Skills: Emphasize debugging challenges (e.g., numpy incompatibility, dataset limitations) and solutions, showcasing technical resilience.
3. Business Impact: Highlight applications in media monitoring and business intelligence, using visualizations (e.g., topic distribution plots) and metrics (e.g., accuracy, F1-score) to appeal to employers in tech or media industries.

We will include the notebook, a demo video running the pipeline on sample articles, and a slide deck summarizing results and business applications, tailored for roles in NLP, data science, or AI product development.

Professional Development Impact

This project has significantly shaped our professional growth:

1. Technical Expertise: Mastering NLP pipeline integration and debugging complex errors (e.g., LDA dependencies) has prepared us for data science roles requiring robust system design.
2. Leadership Skills: Leading the team through coordination challenges honed project management and communication abilities, critical for collaborative tech environments.
3. Business Acumen: Translating NLP outputs into business insights (e.g., sentiment for crisis management) has deepened our understanding of AI’s commercial value.
4. Ethical Awareness: Addressing biases and privacy concerns aligns with our goal of building responsible AI systems.

This experience has solidified the passion for NLP and motivated to pursue advanced topics like transformers and real-time analytics, positioning us for impactful contributions in AI-driven industries.

Conclusion

The NewsBot Intelligence System project was a challenging yet rewarding opportunity to integrate NLP techniques into a practical, industry-relevant system. Overcoming technical hurdles like dependency errors and dataset limitations strengthened coding and problem-solving skills, while the LDA Research Extension added innovative depth. The project's real-world applications in media monitoring and business intelligence underscored its value, and ethical reflections emphasized the need for fairness in NLP. Despite team collaboration challenges, leading the project and delivering a robust system with Sha'Rise's support was a proud achievement. This project has equipped us with the skills and confidence to tackle complex NLP challenges and inspired us to explore advanced techniques in future endeavours.