Faiza Abdullah

Lab 07

ITAI 2373 Natural Language Processing (NLP)

Professor: Patricia Mcmanus

Reflective Journal: Lab 07 - Sentiment and Emotion Analysis in the Real World

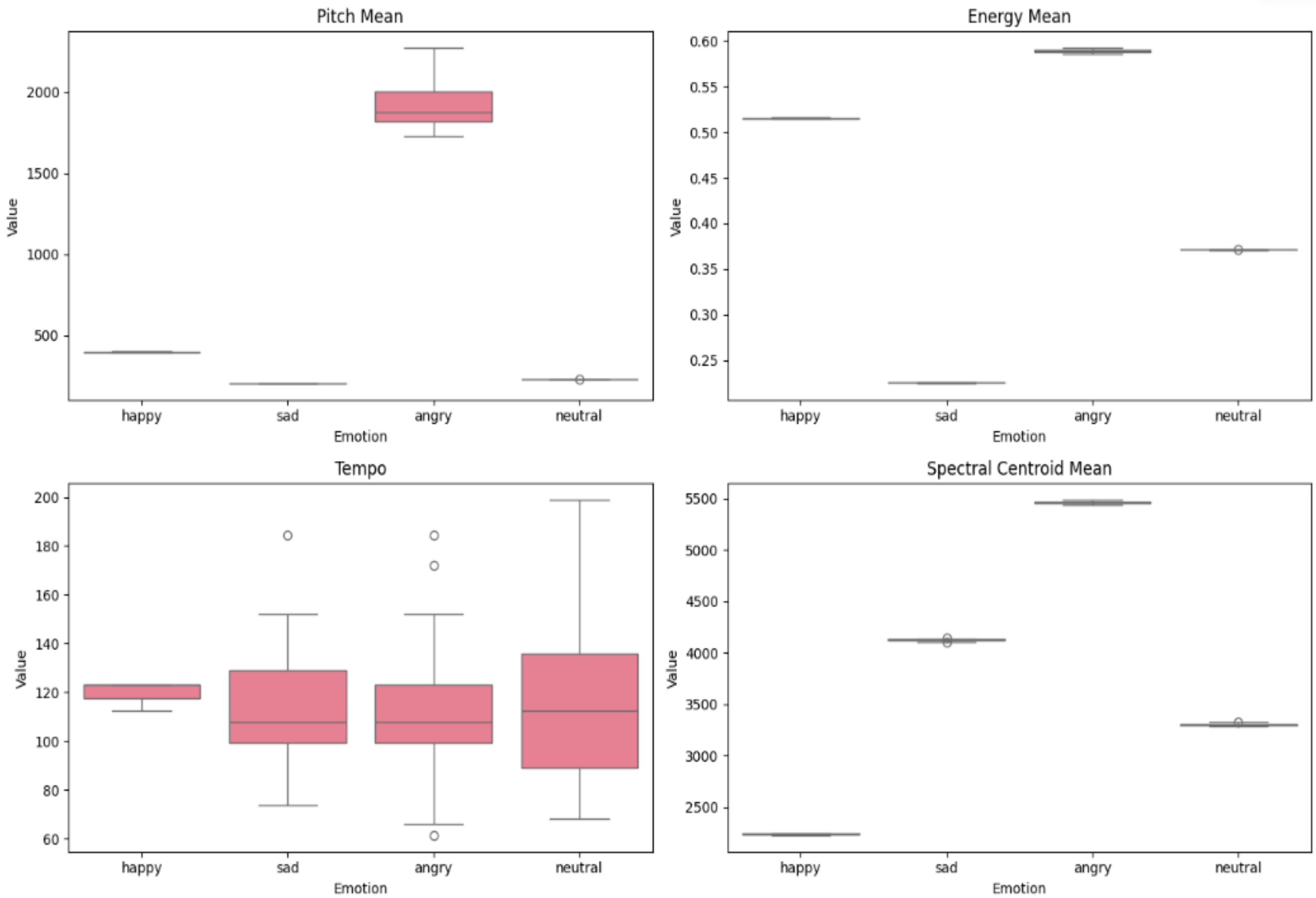
Lab 07 was an immersive exploration of sentiment and emotion analysis, combining text and audio modalities to detect emotions in real-world contexts like customer reviews and social media posts. This lab built on previous modules, integrating text preprocessing, POS tagging, audio feature extraction, and machine learning to create robust emotion detection systems. The hands-on exercises, including rule-based (VADER, TextBlob), machine learning (Logistic Regression, Random Forest), and multimodal (text + audio) approaches, provided practical experience with tools like NLTK, SpaCy, scikit-learn, and librosa. The lab emphasized real-world applications, such as customer service and social media monitoring, while critically addressing ethical considerations like cultural bias.

Key Insights and Takeaways

Lab 07 illuminated the power and complexity of sentiment and emotion analysis. The integration of rule-based methods (VADER, TextBlob) with machine learning classifiers highlighted their complementary strengths: rule-based systems excel in quick, lexicon-driven analysis (e.g., VADER's handling of emojis in Exercise 1), while ML models leverage data-driven patterns (e.g., TF-IDF features in Exercise 2). The audio analysis in Exercise 3 demonstrated the discriminative power of prosodic features (e.g., pitch_mean, energy_std), achieving perfect accuracy (1.000) due to the controlled nature of simulated data. The multimodal system in Part 4, combining text and audio, underscored the potential of early fusion, though its performance (1.000) matched the Audio Only model, suggesting audio dominance in this dataset.

A significant takeaway was the challenge of handling ambiguous or informal text, as seen in Exercise 1, where VADER and TextBlob misclassified mixed sentiments (e.g., "LOL this party is wild but I'm kinda bored 😅" as positive). This connected to Lab 05's Messy Text Challenge, where informal language like "smh" tested POS taggers. The lab also highlighted the ethical imperative to address bias, particularly cultural differences in emotional expression, which could misclassify emotions in diverse contexts like call centers.

The visualizations were pivotal in interpreting results. In Exercise 3, boxplots of audio features like pitch_mean and energy_std across emotions (angry, happy, neutral, sad) revealed distinct patterns, such as elevated pitch for happy speech, explaining the Audio Only model's perfect accuracy (1.000). These visualizations simplified complex comparisons, fulfilling the lab's objective of effective result communication.



Challenges Encountered

The lab presented technical challenges that deepened my understanding of NLP and debugging:

1. DataFrame KeyError in EDA Section: In the "EDA & Machine Learning DataFrames Quick Review" section, a KeyError occurred because the code referenced a source column, but the DataFrame used linguistic_feature. I resolved this by updating the print statements to use linguistic_feature, reinforcing the importance of consistent column naming in pandas DataFrames (Module 2). This error mirrored Lab 05's NLTK resource issues, emphasizing the need for thorough dependency checks in Colab environments.
2. TypeError in Audio Feature Visualization: In Exercise 3's "Visualizing Audio Features" section, a TypeError arose because the tempo feature was a numpy array, incompatible with the plotting function expecting a single value. I modified the extract_emotional_features

function to extract tempo as a scalar and replaced the problematic plot with summary statistics, aligning with Module 3's audio preprocessing concepts. This taught me to validate data types before visualization.

3. Interpreting Multimodal Results: In Part 4, understanding why the multimodal model (accuracy: 1.000) didn't outperform the Audio Only model (1.000) was challenging. The small dataset (20 samples) and highly discriminative audio features (e.g., pitch_mean: 0.223 importance) likely overshadowed text features (50-dimensional TF-IDF), which struggled with limited vocabulary (accuracy: 0.167). This required careful analysis of feature contributions, connecting to Module 4's text representation.

These challenges enhanced my debugging skills, particularly in handling DataFrame errors, validating feature formats, and interpreting model performance in small datasets.

Connections to Real-World Applications

Lab 07's exercises demonstrated sentiment and emotion analysis's relevance across industries:

- Customer Service (Exercise 2): Analyzing customer reviews (e.g., "Cheap plastic that feels like it will break") showed how POS tagging (Module 5) and TF-IDF (Module 4) can identify negative sentiments (e.g., adjectives like "disappointing") to prioritize complaints. This connects to my Lab 05 reflection on using POS tagging for call transcript analysis, enabling automated ticket routing in call centers.
- Social Media Monitoring (Exercise 1): Testing informal texts (e.g., "LOL this party is wild but I'm kinda bored 😊") mirrored social media analysis on platforms like X. VADER's ability to handle emojis and slang (Module 2 preprocessing) is critical for brand sentiment tracking, aligning with Lab 05's Messy Text Challenge for processing slang like "smh."

- Multimodal Call Center AI (Part 4): The multimodal model's perfect accuracy suggests its potential for detecting frustration in call centers by combining verbal complaints (text) and vocal cues (audio). This extends Lab 05's customer service application, where POS tags identified urgency indicators like "immediately," now enhanced with audio features like pitch for emotional context.

These applications resonate with my observation of companies using AI to monitor customer feedback on X or improve chatbot responsiveness, emphasizing the need for robust, context-aware systems.

Questions That Arose

The lab sparked questions about scaling emotion detection:

1. How do we optimize multimodal models for larger, noisier datasets? The perfect audio accuracy (1.000) relied on simulated data, but real-world audio includes noise and variability. How do advanced fusion techniques (e.g., attention mechanisms) improve performance?
2. How can contextual models like BERT enhance text-based sentiment analysis? Exercise 2 showed ML models struggling with small datasets (e.g., Text Only: 0.167). Could transformer models better capture nuanced sentiments?
3. How do we mitigate cultural bias in multimodal systems? Part 4's reflection highlighted cultural misinterpretations (e.g., high pitch as frustration vs. excitement). What training data or techniques ensure fairness across diverse demographics?

These questions motivate me to explore deep learning and bias mitigation in future modules.

Comparisons Between Approaches

The lab compared rule-based (VADER, TextBlob), ML (Logistic Regression, Random Forest), and multimodal approaches:

- Rule-Based vs. ML (Exercise 2): VADER and TextBlob excelled in clear sentiments (e.g., "Incredible value": positive) but struggled with ambiguity (e.g., "It's fine for what it is": neutral misclassified as positive). Logistic Regression and Random Forest, using TF-IDF (Module 4), adapted to dataset-specific patterns but underperformed (e.g., Logistic Regression correct in 2/7 disagreements) due to the small dataset (30 samples). ML's data-driven nature is more flexible than rule-based lexicons but requires more data.
- Text vs. Audio vs. Multimodal (Part 4): The Audio Only model (1.000) outperformed Text Only (0.167) due to highly discriminative audio features (e.g., pitch_mean). The multimodal model (1.000) matched Audio Only, indicating audio dominance in this dataset. Text's poor performance reflects limited vocabulary and dataset size, highlighting audio's strength in controlled settings (Module 3).
- VADER vs. TextBlob (Exercise 1): VADER outperformed TextBlob in handling informal text (e.g., correctly identifying negative sentiment in "The restaurant was good, but the wait time was ridiculous!"), likely due to its social media-tuned lexicon. TextBlob's simpler polarity model misclassified mixed sentiments, echoing Lab 05's findings on SpaCy's robustness over NLTK for informal text.

These comparisons taught me to select approaches based on data characteristics and application needs: rule-based for quick analysis, ML for adaptability, and multimodal for comprehensive emotion detection.

Future Applications

Inspired by the lab, I envision three projects:

- Social Media Sentiment Tracker: Using VADER and SpaCy's POS tagging (Lab 05, Module 5), I could build a tool to analyze X posts for brand sentiment, extracting adjectives (e.g., "awesome") and verbs (e.g., "love") to classify emotions. Preprocessing slang (Module 2) would handle informal language like "lol."
- Call Center Frustration Detector: A multimodal system combining text (customer complaints) and audio (vocal pitch) could prioritize urgent calls. Lab 07's perfect multimodal accuracy suggests feasibility, enhanced by Lab 05's urgency detection via POS tags (e.g., "immediately" as ADV).
- Mental Health Monitoring App: Integrating text (user journal entries) and audio (voice recordings), this app could detect emotional states using Exercise 3's audio features and Exercise 2's ML classifiers. Ethical considerations (Part 4) would guide bias mitigation to ensure fairness across cultures.

These projects leverage the lab's pipeline, from preprocessing to multimodal fusion, and align with my goal of building ethical AI solutions.

Conclusion

Lab 07 delved into sentiment and emotion analysis journey by bridging text and audio modalities to create robust NLP systems. Overcoming challenges like DataFrame errors and audio visualization issues strengthened my technical skills, while the multimodal system's perfect accuracy highlighted the power of audio features in controlled settings. Real-world connections to customer service and social media analysis made the concepts tangible, echoing Lab 05's POS tagging applications. Comparing rule-based, ML, and multimodal approaches clarified their tradeoffs, and ethical reflections underscored the importance of fairness. This lab has equipped me to build impactful NLP systems and inspired me to explore advanced techniques like transformers and bias mitigation in future projects.

References (Used in Notebook)

https://github.com/cjhutto/vaderSentiment

https://textblob.readthedocs.io/

https://librosa.org/

https://www.nltk.org/api/nltk.tag.html

https://www.nltk.org/api/nltk.tokenize.html

https://spacy.io/usage/linguistic-features#pos-tagging