Project Proposal: A Hybrid Stylometric and Semantic Fluctuation Approach for Detecting AI-Generated Media News

Motivation:

- Large Language Models (LLMs) pose a direct threat to the integrity of public information as vast quantities of convincing but false media news can be generated at scale.
- Existing AI text detectors often rely on content-based classifiers that are vulnerable to adversarial attacks and out-of-distribution data, creating a need for more robust detection methods.
- Recent work shows that contrastive learning and multi-level analysis can improve detection robustness (Chen et al., 2024).

Research Question: Can a hybrid model combining statistical stylometry with semantic fluctuation analysis achieve greater accuracy and robustness in detecting AI-generated news compared to standard content-based classifiers, particularly in out-of-distribution scenarios?

Proposed Approach:

1. **Dataset Construction:**
    - Scrape human-written news sources (Reuters, AP, BBC) as human baseline
    - Generate synthetic counterparts using multiple LLMs (GPT-4o, Claude, Gemini) on same topics
    - Create adversarial dataset by having humans paraphrase AI-generated articles to simulate evasion tactics
2. **Feature Engineering - Stylometric Analysis:**
    - Lexical diversity metrics (type-token ratio, hapax legomena)
    - Sentence complexity patterns (parse tree depth, dependency distances)
    - N-gram frequency distributions
    - Punctuation and formatting patterns
3. **Feature Engineering - Semantic Fluctuation Analysis (Novel Contribution):**
    - **Topic Coherence Drift:** Measure semantic similarity between consecutive paragraphs using sentence embeddings, tracking unusual coherence patterns
    - **Entity Consistency Score:** Analyze how consistently named entities and their attributes are referenced throughout the article
    - **Temporal Logic Patterns:** Detect inconsistencies in temporal references and event sequencing
    - **Source Attribution Patterns:** Analyze frequency and specificity of source citations and quotes
    - **Factual Grounding Metrics:** Measure the density and specificity of verifiable claims vs. vague statements
4. **Model Development:**
    - Implement ensemble classifier combining stylometric and semantic features
    - Use contrastive learning approach inspired by DeTeCtive framework for robust feature extraction

    - Apply multi-task learning to simultaneously detect AI content and identify source model

5. **Evaluation:**
    - Benchmark against existing detectors (GPTZero, OpenAI's classifier)
    - Test on out-of-distribution data from new LLMs
    - Evaluate robustness against adversarial paraphrasing

Timeline (6 weeks):

Week 1: Data Collection & Preparation

- Week 1: Set up web scraping infrastructure, collect 1,000 human-written articles; Generate AI counterparts using 3 different LLMs, create initial dataset splits

Week 2: Feature Development

- Week 2: Implement stylometric feature extractors, validate on sample data; Develop semantic fluctuation analyzers, test feature discriminative power

Week 3-4: Model Training & Optimization

- Week 3: Train baseline models, implement ensemble approach
- Week 4: Hyperparameter tuning, implement contrastive learning components

Week 5: Adversarial Testing

- Create human-paraphrased adversarial dataset
- Test model robustness and identify failure modes

Week 6: Evaluation & Documentation

- Comprehensive benchmarking against existing methods
- Statistical analysis of results
- Final report preparation

Expected Outcomes:

- Demonstrated improvement over existing methods on adversarial examples
- Open-source toolkit for AI-generated news detection
- Analysis of most discriminative features for detection

References:

Chen, J., et al. (2024). "DeTeCtive: Detecting AI-generated Text via Multi-Level Contrastive Learning." *arXiv preprint arXiv:2410.20964*.

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). "GLTR: Statistical Detection and Visualization of Generated Text." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 111-116.

Ippolito, D., et al. (2020). "Automatic Detection of Generated Text is Easiest when Humans are Fooled." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1808-1822.

Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. V. (2020). "Automatic Detection of Machine Generated Text: A Critical Survey." *Proceedings of the 28th International Conference on Computational Linguistics*, 2296-2309.

Mitchell, E., et al. (2023). "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature." *Proceedings of the 40th International Conference on Machine Learning*, 24950-24962.

Sadasivan, V. S., et al. (2024). "Can AI-Generated Text be Reliably Detected?" *arXiv preprint arXiv:2303.11156*.

Zellers, R., et al. (2019). "Defending Against Neural Fake News." *Advances in Neural Information Processing Systems*, 32.