Research & Development Plan: Sentiment Analysis Model Selection

To: Person A (Model Researcher)

From: Gemini AI

Date: 04 August 2025

Subject: A Step-by-Step Plan for Selecting a Foundational Sentiment Analysis Model

1.0 Objective

The primary objective of this plan is to systematically identify, evaluate, and select the most suitable pre-trained sentiment analysis model for integration into our financial news-based trading bot. The final selection must be data-driven, well-documented, and based on both theoretical soundness and practical feasibility.

2.0 Research Phases & Actionable Steps

This project is divided into four distinct phases. Each phase has specific tasks and a clear deliverable required to proceed to the next phase.

Phase 1: Landscape Analysis & Candidate Identification (Duration: 2-3 Days)

Goal: To map the current landscape of financial sentiment analysis and create a shortlist of 2-3 primary candidate models.

- **Task 1.1: Academic Literature Review**
  - **Action:** Utilize academic search engines (Google Scholar, arXiv.org, IEEE Xplore) to find survey papers and highly-cited articles on financial sentiment analysis.
  - **Search Queries:** Use terms such as "financial sentiment analysis survey", "stock prediction using NLP", "FinBERT", "transformer models for finance".
  - **Focus On:** Identifying the models and techniques that are consistently cited as state-of-the-art. Pay attention to the datasets used for benchmarking.
- **Task 1.2: Practitioner & Open-Source Review**
  - **Action:** Investigate platforms where these models are used in practice.
  - **Platforms:**
    - **Hugging Face:** Search the Models hub for "finance", "sentiment". Filter by most downloads/likes to identify community-preferred models. Read the model cards carefully.
    - **GitHub:** Search for open-source projects using the models identified in

Task 1.1. Look for well-maintained repositories with clear documentation.
        - **Kaggle:** Review past competitions related to stock market prediction or sentiment analysis. Analyze the methodologies of the winning solutions.
- **Phase 1 Deliverable:**
    - A concise research brief (1-2 pages). This document must contain:
        1. A summary of the current state-of-the-art based on academic literature.
        2. A shortlist of the top 3 candidate models, with a justification for each based on community adoption, documentation quality, and initial relevance to our project.

Phase 2: Deep Dive Analysis of Primary Candidates (Duration: 2-3 Days)

Goal: To gain a deep technical understanding of the shortlisted models.

- **Task 2.1: Foundational Paper Analysis**
    - **Action:** For each of the top 2-3 candidates, locate and perform a deep reading of the original academic paper.
    - **Focus On:**
        - **Architecture:** How is the model built?
        - **Training Data:** What specific corpus was it trained and fine-tuned on? (e.g., Financial PhraseBank, news articles, etc.). This is crucial for understanding its potential biases.
        - **Reported Performance:** What were the reported accuracy, precision, and recall scores?
        - **Limitations:** What weaknesses or limitations did the authors explicitly mention?
- **Task 2.2: Implementation & Dependency Review**
    - **Action:** Find the canonical implementation of each candidate model (likely on Hugging Face or a dedicated GitHub repository).
    - **Focus On:**
        - **Input/Output:** What is the exact format of the input data (string, list of strings)? What is the format of the output (class labels, probabilities, dictionary)?
        - **Dependencies:** What libraries are required (transformers, PyTorch, TensorFlow)? Note the specific versions.
- **Phase 2 Deliverable:**
    - A technical evaluation document for each candidate model. This should include a summary of the academic paper and a section detailing the practical implementation requirements.

Phase 3: Empirical Validation & Prototyping (Duration: 2 Days)

Goal: To test the leading candidate model with our own data in a controlled environment.

- **Task 3.1: Sandbox Environment Construction**
  - **Action:** Create a new, isolated Python environment (using conda or venv) for the leading candidate model.
  - **Action:** Install all necessary dependencies identified in Task 2.2. Create a requirements.txt file.
- **Task 3.2: Feasibility Test**
  - **Action:** Write a Python script to perform a feasibility test.
  - **Script Logic:**
    1. Load a sample of 200-300 headlines from your headlines_20250426.csv file.
    2. Instantiate the pre-trained model and its tokenizer.
    3. Loop through the sample headlines, passing each one to the model for sentiment prediction.
    4. Store the results (headline, predicted label, confidence score) in a new CSV file.
- **Phase 3 Deliverable:**
  1. The fully documented Python script used for the feasibility test.
  2. The output CSV file containing the model's predictions on our dataset.

Phase 4: Synthesis & Final Recommendation (Duration: 1 Day)

Goal: To make a final, evidence-based decision and prepare for integration.

- **Task 4.1: Qualitative Results Analysis**
  - **Action:** Manually review the output CSV from the feasibility test.
  - **Focus On:** Where does the model perform well? Where does it fail? Are there any systematic errors (e.g., misinterpreting financial jargon)? How does it handle ambiguous headlines?
- **Task 4.2: Formulate Recommendation**
  - **Action:** Based on the entire body of research, formulate a final recommendation.
  - **Action:** Prepare a concise presentation or report to deliver to Person B.
- **Phase 4 Deliverable:**
  - A final recommendation presentation. It must include:
    1. The chosen model and a clear, evidence-based justification for the choice.
    2. A summary of the model's strengths and acknowledged limitations.
    3. The working prototype script and the results of the feasibility test.

4. A clear statement on readiness for integration into the main trading bot application.