Northeastern University, Khoury College of Computer Science

CS 6120 Natural Language Processing
CS6120 Final Project Proposal

Due: Dec 2025(100 points)

Soonbee Hwang, Xinyuan Fan (Amber)
hwang.soon@northeastern.edu, fan.xinyua@northeastern.edu

1 Proposed Objective

We propose to build a Financial Market Intelligence Assistant powered by Retrieval-Augmented Generation (RAG). The system retrieves relevant daily news headlines from a large-scale financial dataset and generates grounded, citation-backed answers using a local LLM. The system addresses hallucinations, provides verifiable evidence, and demonstrates how external domain knowledge can enhance LLM reasoning in finance.

2 Motivation and Impact

Financial analysis demands:

- high factual accuracy,
- up-to-date information,
- multi-document reasoning, and
- verifiable claims.

Large Language Models (LLMs), however, suffer from knowledge cutoffs and hallucinations, which make them unreliable for real-time financial use. RAG provides a solution by forcing models to reference retrieved documents, enabling transparency and reducing hallucination rates.

This project aims to demonstrate how modern RAG architectures can significantly strengthen financial analysis for investors, students, analysts, and researchers. The impact includes more trustworthy insights, improved market understanding, and automated retrieval of relevant, high-quality financial evidence.

3 Background, Relevant Work, and Dataset

Financial information changes rapidly, and conventional LLMs cannot independently access or reason over the latest data. RAG systems, widely validated in industry and academia, have shown strong performance in domains requiring factual reliability. Our work builds on existing literature on:

- context from multiple news sources,
- accurate historical information,
- interpreting macroeconomic signals, and
- avoiding hallucinated facts.

Datasets

We use the Kaggle "Daily News for Stock Market Prediction" dataset, which provides more than 50,000 financial news headlines (Top 25 headlines per day from 2008–2016). This dataset is ideal for RAG because it consists of:

- News data from Reddit WorldNews Channel
- Stock data (Downloaded directly from Yahoo Finance)

These sources reflect real-world financial decision processes and provide rich, multi-document evidence for retrieval.

4 Proposed Approach / Implementation Details

Our pipeline integrates data processing, retrieval, ranking, and grounded generation. Key methodological components include:

Pre-processing

- Clean and normalize text
- Lowercasing and removing punctuation
- Deduplicating repeated headlines

Retrieval Backend (TF-IDF RAG)

To reduce complexity while maintaining effectiveness, our primary retrieval method uses:

- using semantic chunking to prevent overweighting long documents,
- applying retrieval filtering (Top-K + score thresholding),
- optionally adding reranking in future work.

Learning Algorithms

- Embeddings: BGE-Large-en for high-quality semantic vectors
- Vector Store: FAISS IndexFlatL2
- LLM: local Mistral 7B GGUF via llama-cpp-python
- RAG: retrieve → augment prompt → generate answer with citations

Validation and Evaluation

We will evaluate the system based on:

- retrieval relevance and semantic similarity,
- correctness of citations,
- end-to-end latency (target < 1.5 seconds),
- clarity of the user interface,
- robustness to noisy or ambiguous queries.

User Interface

A Streamlit-based UI including:

- query input panel,
- generated answer with citations,
- adjustable retrieval parameters,
- latency display,
- expandable evidence view.

Deployment

We will Dockerize the entire system and host it on a GCP VM for Demo Day.