Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models

A short Story Presentation

By Aditi Patil

Introduction

Key Focus:

- Financial sentiment analysis: Essential for interpreting emotions in financial texts for market forecasting and investor insights.
- Challenges with traditional NLP and LLMs: Limited understanding of complex financial content and generalization issues.

Large Language Models (LLMs) in Finance:

- Superior performance yet facing challenges in precise sentiment prediction due to training-objective discrepancy and lack of contextual depth.

Our Innovative Approach:

- Introducing a Retrieval-Augmented Large Language Model Framework.
- Combines instruction-finetuned LLMs with contextual enrichment from external sources.
- Achieves 15% to 48% better accuracy and F1 score in financial sentiment analysis compared to existing models.

Background

A. Financial Sentiment Analysis

- Early Research: Focused on fine-tuning pre-trained models. Limited in complex scenarios, especially with numerical data.
- LLMs Emergence: Increasing model size and data enhanced in-context learning and zero-shot predictions.

B. Instruction Tuning in LLMs

- Training Techniques: Causal Language Modeling leading to randomness in outputs.
- Solution: Instruction tuning for precise, user-directed output generation.

C. Retrieval Augmented Generation (RAG)

- Methodology: Combines context retrieval and LLMs for language generation.
- Dual Knowledge Source: Uses both parametric memory (LLMs) and nonparametric memory (external documents).

Approach used

Module 1 :Instruction-tuned LLM Module

- Objective: Align LLM behavior with financial sentiment prediction.
- Process: Fine-tune an open-source LLM (e.g., LLaMA, ChatGLM) using a custom instruction-following dataset specific to financial sentiment.

Module 2: Retrieval Augmented Generation (RAG) Module

- Function: Retrieves relevant background information from various external sources.
- Sources: Includes Bloomberg, Reuters, Goldman Sachs, Citi Velocity, Twitter, and Reddit.
- Method: Uses a multi-source query and similarity-based retrieval for context enrichment.

Integration and Outcome

- Combining Modules: The retrieved context is merged with the original query to form the final query.
- Sentiment Prediction: The instruction-tuned LLM generates sentiment predictions based on this enriched query, enhancing prediction accuracy.

Instruction Tuned LLMs

Overview of Instruction Tuning

- Purpose: Align LLMs with financial sentiment prediction tasks.
- Effectiveness: Proven to adhere well to user instructions in predicting financial sentiments.

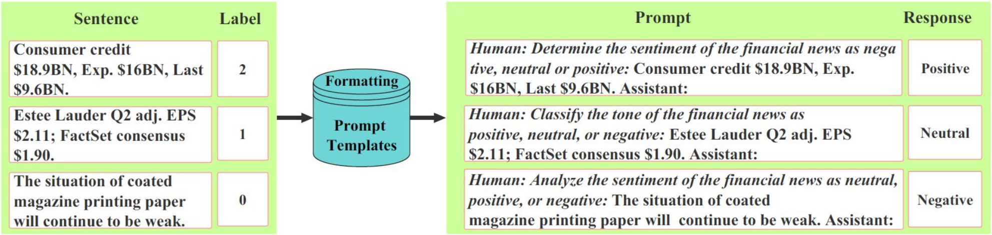
Process of Instruction Tuning

1. **Constructing the Dataset:**Create an instruction-following dataset with paired instructions and expected sentiment responses (positive, negative, neutral).
2. **Fine-Tuning LLMs:** Utilize the dataset for fine-tuning, teaching LLMs to generate accurate responses based on financial sentiment instructions.
3. **Mapping Outputs to Sentiment Classes:** Convert the model's generative outputs back into predefined sentiment classes for proper evaluation.

Instruction Tuned LLMs

Detailed Steps

- Dataset Creation:
  - Convert existing financial sentiment datasets into instruction-following format. Include human-written instructions and corresponding outputs.
- Fine-Tuning Methodology:
  - Tokenize texts using Byte-Pair Encoding (BPE).
  - Apply Causal Language Modeling (CLM) to optimize prediction of next tokens.
- Output Mapping:
  - Sequentially check generated text for sentiment indicators ("negative", "neutral", "positive") and categorize accordingly.



Sentence	Label
Consumer credit $18.9BN, Exp. $16BN, Last $9.6BN.	2
Estee Lauder Q2 adj. EPS $2.11; FactSet consensus $1.90.	1
The situation of coated magazine printing paper will continue to be weak.	0

Prompt	Response
Human: Determine the sentiment of the financial news as negative, neutral or positive: Consumer credit $18.9BN, Exp. $16BN, Last $9.6BN. Assistant:	Positive
Human: Classify the tone of the financial news as positive, neutral, or negative: Estee Lauder Q2 adj. EPS $2.11; FactSet consensus $1.90. Assistant:	Neutral
Human: Analyze the sentiment of the financial news as neutral, positive, or negative: The situation of coated magazine printing paper will continue to be weak. Assistant:	Negative

Fig. 2: Formatting sentiment analysis dataset into instruction-following dataset.

RAG Module

Two-Step Knowledge Retrieval Process

Multi-Source Knowledge Query:

- Preprocess text to remove irrelevant content.
- Utilize retrieval APIs to extract information relevant to the financial query.

Similarity-Based Retrieval:

- Use the Szymkiewicz-Simpson coefficient for higher relevance.
- Focus on exact matches and minimize irrelevant retrievals.

Integration with Instruction-tuned LLM

- Combine retrieved context with original input to form enriched data for LLM response generation.

The Retrieval Augmented Generation (RAG) Module

Overview of RAG

- Purpose: Enhances LLM accuracy by injecting external knowledge into response generation.
- Process: Involves setting up external knowledge sources, two-step knowledge retrieval, and integration with input data.

Setting Up External Knowledge Sources

- News Sources: Bloomberg, Yahoo Finance, Reuters, CNBC, Market Screener for consistent, reliable information.
- Research Publications: Centralized (Goldman Sachs, Citi) and crowd-based (Seeking Alpha) platforms for diverse financial insights.
- Social Media: Twitter, Reddit for real-time updates; requires careful analysis due to volatility.

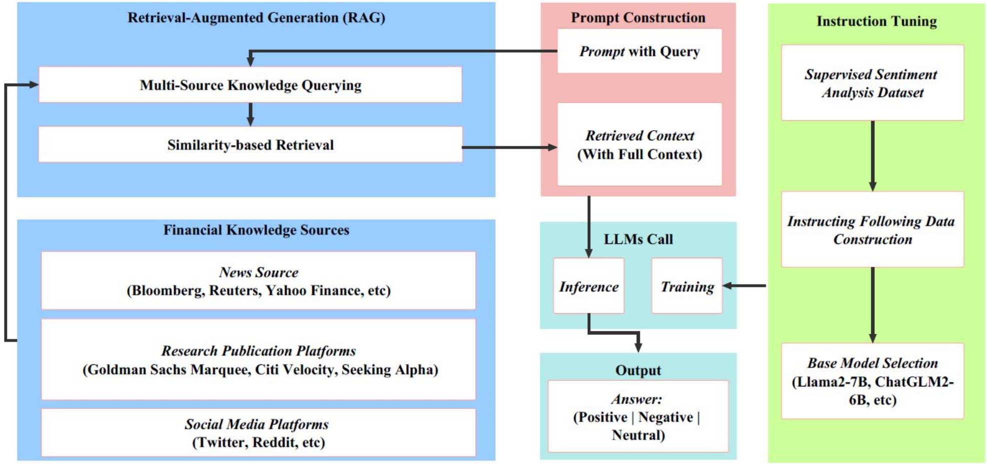

Retrieval-Augmented Generation (RAG)
Multi-Source Knowledge Querying
Similarity-based Retrieval
Financial Knowledge Sources
News Source
(Bloomberg, Reuters, Yahoo Finance, etc)
Research Publication Platforms
(Goldman Sachs Marquee, Citi Velocity, Seeking Alpha)
Social Media Platforms
(Twitter, Reddit, etc)
Prompt Construction
Prompt with Query
Retrieved Context
(With Full Context)
LLMs Call
Inference
Training
Output
Answer:
(Positive | Negative | Neutral)
Instruction Tuning
Supervised Sentiment Analysis Dataset
Instructing Following Data Construction
Base Model Selection
(Llama2-7B, ChatGLM2-6B, etc)

Fig. 1: Framework of retrieval-augmented large language model for financial sentiment analysis.

Why Szymkiewicz-Simpson Coefficient?

Why Szymkiewicz-Simpson Over Semantic Similarity?

Focus on Exact Matches:

- Essential for financial terms like tickers.
- Ensures precision in matching specific financial terms, crucial for accurate sentiment analysis.

Effectively Manages Short-to-Long Text Matching:

- Addresses the challenge of aligning short financial texts (tweets, headlines) with longer contextual documents.
- Prevents the overshadowing of relevant content by document length.

Szymkiewicz-Simpson Coefficient in Detail:

Overlap Measure:

- Calculates similarity by the proportion of overlapping words between input and context.

$$\mathrm{overlap}(\mathbf{X}, \mathbf{Y}) = \frac{|\mathbf{X} \cap \mathbf{Y}|}{min(|\mathbf{X}|, |\mathbf{Y}|)},$$

Advantages in Financial Analysis:

- Prioritizes hard matching for crucial financial terms.
- More effective in specific term detection compared to semantic similarity.

Datasets used

Training Datasets

- **Twitter Financial News Dataset:**
  - Corpus: News tweets related to the financial sector.
  - Purpose: Classify financial sentiment within Twitter discussions.
  - Size: 9,540 samples.
  - Labels: Bearish, Bullish, Neutral.
- **FiQA Dataset:**
  - Inclusion: Diverse financial texts.
  - Size: 961 samples.
  - Labels: Positive, Neutral, Negative.

Testing Datasets

- **Twitter Financial News Sentiment Validation (Twitter Val):**
  - Split: Validation part of the Twitter dataset.
  - Size: 2,388 samples.
  - Challenge: Often lacks clear sources and context.
- **Financial PhraseBank (FPB) Dataset:**
  - Source: Financial news articles from Lexis-Nexis.
  - Size: 4,840 samples.
  - Annotation: By 16 finance and business professionals for high-quality sentiment labels.

Evaluations and Analysis

Overview of Evaluation

- Objective: Assess effectiveness of instruction tuning and RAG.
- Method: Compare against state-of-the-art sentiment analysis models and general-purpose LLMs.

Training and Testing Datasets

- Training Data: Combination of Twitter Financial News and FiQA datasets (10,501 samples).
- Testing Data: Twitter Val and Financial PhraseBank (FPB) datasets.

Model Training

- Base Model: Llama-7B.
- Training Process: 10 epochs, AdamW optimizer, batch size of 32, max input text length of 512 tokens.
- Hardware: DeepSpeed on 8×A100 GPUs, 58 minutes total training time.

Baseline Models for Comparison

- Models: BloombergGPT, ChatGPT, Llama-7B, ChatGLM2-6B, FinBERT.
- Performance Metrics: Accuracy and F1-score.

Performance Results

- Our Model's Performance: Significantly outperforms baselines in accuracy and F1 score.
- Instruction Tuning Impact: Demonstrates improved ability to discern sentiment.

Case Study of RAG Module

Assessing RAG's Effectiveness

- Tested On: Instruction-tuned model and ChatGPT 4.0 using the Twitter Val dataset.
- Results: Introduction of RAG enhances LLMs' performance, validating the impact of added context.

Comparative Performance

- Without RAG:
  - ChatGPT 4.0: 78.8% Accuracy, 65.2% F1 score.
  - Our Model: 86.3% Accuracy, 81.1% F1 score.
- With RAG:
  - ChatGPT 4.0: 81.3% Accuracy, 70.8% F1 score.
  - Our Model: 88.1% Accuracy, 84.2% F1 score.

Case Study Highlight

- Example: $ENR - Energizer shakes off JPMorgan’s bear call.
- Without RAG: Classified as "Neutral".
- With RAG: Context from Seeking Alpha clarifies the phrase, reclassified as "Positive".
- Significance: Showcases how RAG enhances comprehension and nuanced understanding of financial sentiment.

Conclusion

1. **Conclusion**
   - Innovative Framework: Developed a retrieval-augmented LLM framework for enhanced financial sentiment analysis.
   - Instruction Tuning Method: Successfully realigned LLMs for improved accuracy in predicting financial sentiments.
   - Contextual Enrichment: Integrated external knowledge retrieval for more nuanced predictions.
2. **Limitation and Future Work**
   - Current Limitation: Reliance on textual similarity for data retrieval, potentially missing key economic information.
   - Future Direction: Plan to incorporate macroeconomic and microeconomic data to refine analysis accuracy and reliability, aiming for a comprehensive approach in financial sentiment analysis with LLMs.

Thank You !