Assignment#03

REPORT

CS-4045 Deep learning for Perception

Sawab Akbar
I221158
Section A

FAST NUCES, Islamabad

Department of Computer Science

1. Abstract

This study implemented a comprehensive financial text understanding pipeline, starting with Latent Dirichlet Allocation (LDA) to discover latent topics within the dataset. The core component involved a comparative evaluation of three sentiment analysis systems: a **domain-specific FinBERT model**, a **Local LLM (Phi-2) in a zero-shot configuration**, and a **Retrieval-Augmented Generation (RAG) enhanced LLM**. The FinBERT model achieved a superior accuracy of **96.91%** on the full test set, thereby fulfilling the performance requirement and justifying the decision to **skip** conditional fine-tuning.

2. Dataset Description and Preprocessing

The pipeline utilized a dataset of **2,264 financial sentences** labeled with sentiment: 'positive', 'negative', or 'neutral'.

Preprocessing Steps:

1. **Lowercasing**
2. **Punctuation and Number Removal** (using regex [^a-zA-Z\s])
3. **Tokenization** (using nltk.word_tokenize)
4. **Stopword Removal** (using nltk.corpus.stopwords)
5. **Lemmatization** (using nltk.stem.WordNetLemmatizer)

Data Split and Class Distribution

The dataset was split using a stratified approach (80% train, 20% test) to maintain class proportions across subsets.

Sentiment	**Count**
Neutral	1,391
Positive	570
Negative	303

Total Samples	**2,264**

- **Training Set Size**: 1,811 samples
- **Test Set Size**: 453 samples

3. Topic Modeling (LDA) Results

Latent Dirichlet Allocation (LDA) was applied to the preprocessed text tokens using $k=5$ topics to uncover the thematic structure of the financial reports.

LDA Topic Keywords and Coherence

The LDA model achieved a **Coherence Score (C_v) of 0.3898**.

Topic ID	**Top 10 Keywords**	**Interpretation**
0	"mln", "euro", "finnish", "oyj", "helsinki", "share", "said", "stock", "company", "profit"	**Stock Markets & Share Trading** (Focus on European stocks, company announcements, and general profits).
1	"sale", "million", "net", "finnish", "year", "percent", "said", "market", "first", "finland"	**Revenue & Market Performance** (Focus on sales figures, net results, and year-over-year market changes).
2	"company", "share", "total", "said", "value", "cost", "capital", "investment", "eurm", "operation"	**Corporate Strategy & Investment** (Focus on company structure, capital, investment decisions, and operational costs).
3	"eur", "profit", "operating", "quarter", "period", "net", "sale", "decreased", "million", "corresponding"	**Quarterly Financial Results & Decline** (Strong focus on key metrics, quarterly periods, and negative movement keyword "decreased").

4	"service", "market", "expected", "company", "building", "price", "finland", "also", "mobile", "nokia"	**Sector & Product News** (Focus on services, specific industry sectors like mobile, and key companies like Nokia).

4. Sentiment Analysis Comparative Pipeline Design

Three distinct systems were implemented and evaluated to classify the financial sentences into 'positive', 'negative', or 'neutral' sentiments.

A. FinBERT-based Sentiment Analysis

- **Model**: **ProsusAI/finbert** (a BERT model fine-tuned on financial text).
- **Methodology**: Direct batch inference on the full test set (453 samples) using PyTorch, leveraging the domain-specific pre-training.

B. Local LLM (Zero-Shot) Sentiment Analysis

- **Model**: **Phi-2 (microsoft/phi-2)**.
- **Methodology**: **Zero-shot prompting** (no examples provided in the prompt). Due to computational constraints, evaluation was limited to a **50-sample subset** of the test set.

C. RAG-Enhanced LLM (Few-Shot) Sentiment Analysis

- **Model Stack**:
  - **Embeddings**: all-MiniLM-L6-v2 (Sentence Transformer).
  - **Retrieval Index**: **FAISS** (FlatL2 index).
  - **LLM**: **Phi-2 (microsoft/phi-2)**.
- **Methodology**: For each query in the 50-sample subset, the system retrieved the $k=3$ most similar sentences (and their labels) from the training data using FAISS. These examples were then integrated into the LLM prompt, enabling a **few-shot classification** approach.

5. Evaluation and Results

5.1 Performance Metrics Comparison

The table below summarizes the performance metrics for all three systems. FinBERT was evaluated on the full test set, while the LLM and RAG systems were evaluated on a limited 50-sample subset due to inference time.

Method	Accuracy	Precision	Recall	F1-Score	Evaluation Set
FinBERT (Domain Transformer)	**0.9691**	**0.9712**	**0.9691**	**0.9695**	Full Test Set (N=453)
Local LLM (Zero-Shot)	0.6800	0.6913	0.6800	0.6752	Subset (N=50)
RAG-Enhanced LLM (Few-Shot)	0.6000	0.5352	0.6000	0.5069	Subset (N=50)

5.2 Confusion Matrices

The confusion matrices provide a granular view of how each model performed across the three sentiment classes ('positive', 'negative', 'neutral'). The high accuracy of FinBERT is visually evident in the strongly populated diagonals of its matrix, demonstrating effective differentiation between classes.

FinBERT (ProsusAI) Confusion Matrix (Full Test Set)

The FinBERT model achieved superior classification across all classes, showing very few misclassifications.

Local LLM (Zero-Shot) Confusion Matrix (50-Sample Subset)

The Zero-Shot Phi-2 struggled, particularly in correctly identifying the minority classes ('positive' and 'negative'), often misclassifying them as 'neutral'.

RAG-Enhanced LLM Confusion Matrix (50-Sample Subset)

The RAG system performed the poorest on the subset, suggesting that the retrieved examples may have been confusing or noisy, leading to a breakdown in classification logic, particularly affecting the recall for 'positive' and 'negative' sentiments.

5.3 Analysis and Insights

1. **FinBERT Dominance**: The **FinBERT model** exhibited overwhelming superiority, achieving near state-of-the-art performance with **96.91% accuracy**. This strong performance confirms the critical importance of **domain-specific pre-training** in highly

specialized fields like financial text, where sentiment often relies on nuanced vocabulary (e.g., interpreting "raised guidance" as positive).

2. **LLM/RAG Underperformance**: The Phi-2 LLM, even when paired with the RAG system, significantly underperformed FinBERT.
   - The **Local LLM (0.6800 Acc)** struggled in a zero-shot setting, often defaulting to 'neutral' for less obvious positive/negative statements, as the general training corpus lacks deep financial nuance.
   - The **RAG-Enhanced LLM (0.6000 Acc)** surprisingly performed worse than the pure zero-shot LLM. This suggests one or more issues:
     - **Context Quality**: The retrieved $k=3$ examples from the training set might have been misleading or not close enough semantically, confusing the LLM.
     - **Prompt Sensitivity**: The Few-Shot prompt template used for the RAG system may have been less robust than the simpler Zero-Shot prompt.
     - **LLM Latency/Resource Limitations**: The observed discrepancy may also be partially attributed to instability or resource-related issues during complex, multi-step RAG inference on the subset.

6. Conditional Fine-Tuning Justification

As per the assignment constraints, conditional fine-tuning of the FinBERT model was required **only if the maximum achieved accuracy across all three systems was below 90%**.

- **Maximum Achieved Accuracy**: **0.9691** (by FinBERT).
- **Fine-Tuning Decision**: Since $0.9691 \geq 0.90$, the conditional fine-tuning was **skipped**.

The high performance of the pre-trained FinBERT model demonstrates that further training on this specific small dataset would likely yield marginal gains at the cost of significantly increased training time and complexity.

7. Conclusion

The implemented pipeline successfully evaluated modern NLP architectures for financial sentiment analysis. The domain-specific FinBERT model proved to be the most effective solution, providing near-perfect classification due to its specialized knowledge base. While the combination of RAG and a Local LLM offers a flexible path for general NLP tasks, it could not match the performance of a purpose-built transformer in this domain without extensive fine-tuning or a more sophisticated prompt engineering and retrieval strategy. The pipeline is robust, and the results confirmed that no further fine-tuning was necessary to meet high-performance standards.