뉴스 데이터 활용
주식 포트폴리오 최적화

POSTECH IDS 강명구 장승현 홍현종

발표자 : 홍현종

CONTENTS

1 Ploblem Definition

2 Data Collection and Preparation

3 Future Analysis

The Limitations of Quantitative-Only Portfolio Models

Qualitative portfolios may achieve short-term profitability,
but lack long-term reliability and credibility in dynamic market conditions.

PROBLEM DEFINITION

Enhancing Stock Portfolio Reliability using News Sentiment Analysis

Data Input

Market Data
Financial Data

→

Quant Analysis

ML/DL model
MA, Bollinger, ...
Risk analysis
etc

Financial Insight

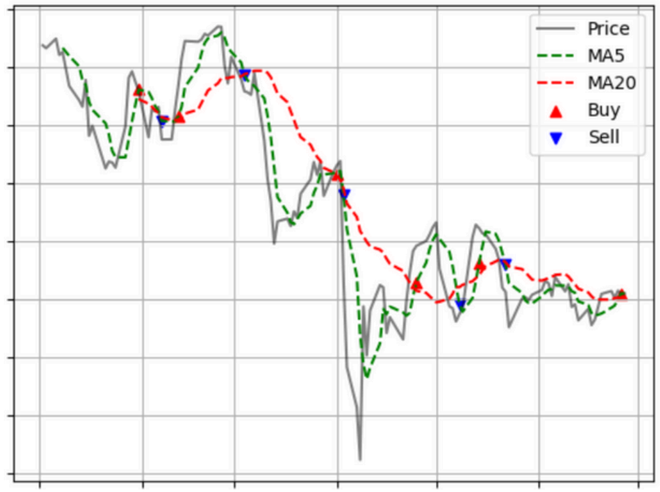
Asset
Liability
Equity

→

Data preprocessing

Grouping / Removing
Normalization
Missing-value handling
Parameter engineering
.
.
.

→



PROBLEM DEFINITION

Enhancing Stock Portfolio Reliability using News Sentiment Analysis

Data Input

Market Data
Financial Data
+ News Sentimental Data

→

Quant Analysis

ML/DL model
MA, Bollinger, ...
Risk analysis
etc

Financial Insight

Asset
Liability
Equity

→

Data preprocessing

Grouping / Removing
Normalization
Missing-value handling
Parameter engineering
.
.
.

→

Sentimental Analysis

$$Index = \frac{poistive - negative}{poistive + negative + 10^{-}6} \times 100$$



Filtering Quantitative Results through News-based Reputation Analysis

Objective in process, **emotional** in understanding

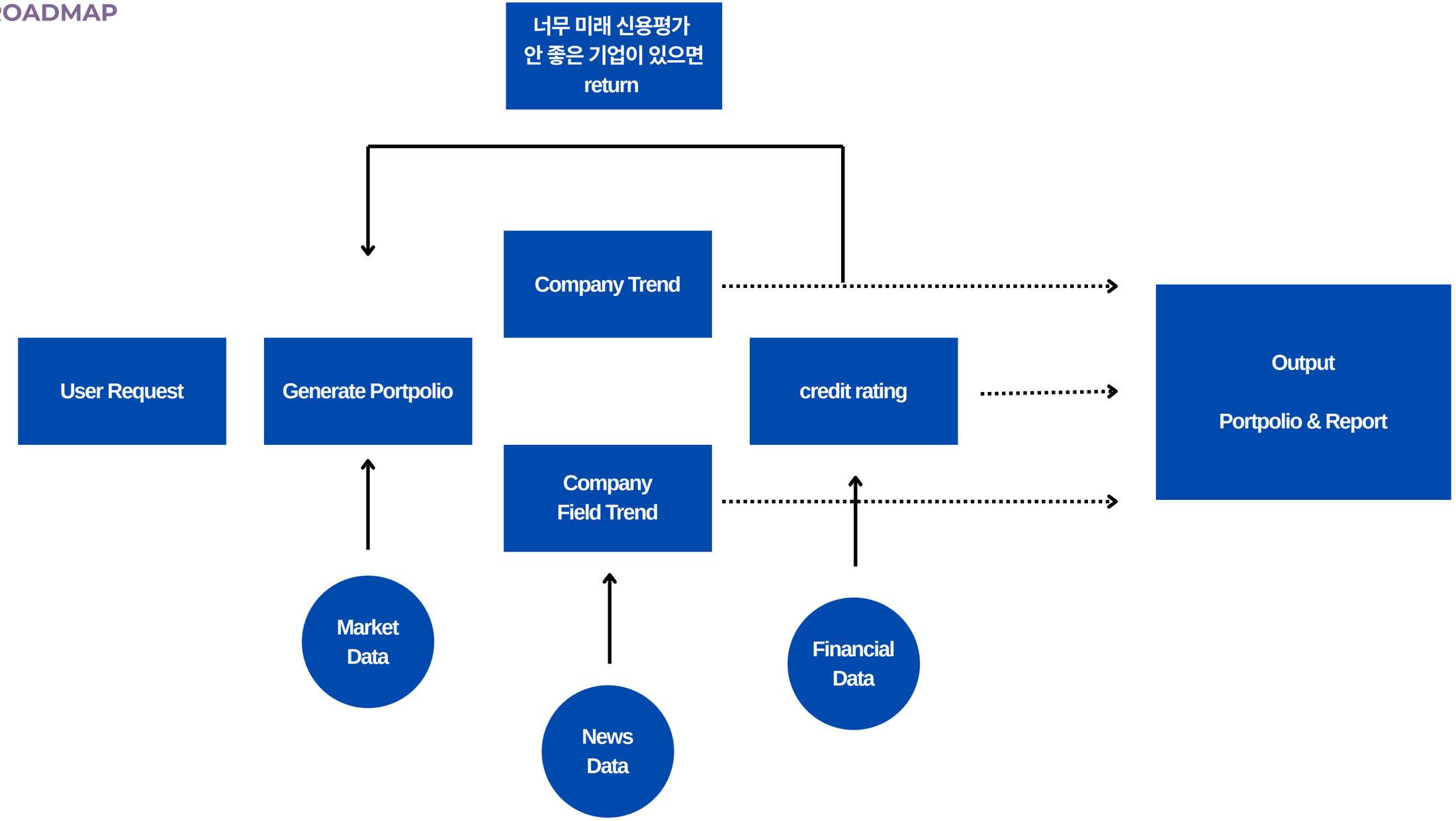
Scope: Domestic equity market

	종목명	현재가	전일비	등락률	액면가	거래량	거래대금	전일거래량	시가	고가	저가	매수호가	매도호가
0	삼성전자	94400.0	상승 5,400	+6.07%	100.0	34625685.0	3250669.0	49883028.0	94000.0	94500.0	92700.0	94300.0	94400.0
1	SK하이닉스	428000.0	상승 32,500	+8.22%	5000.0	6513723.0	2774738.0	7430763.0	426000.0	439250.0	416500.0	427500.0	428000.0
2	LG에너지솔루션	359500.0	하락 39,500	-9.90%	500.0	541859.0	194275.0	644612.0	367500.0	367500.0	352500.0	359000.0	359500.0
3	삼성바이오로직스	1014000.0	상승 5,000	+0.50%	2500.0	73825.0	74881.0	83364.0	1016000.0	1022000.0	1009000.0	1014000.0	1015000.0
4	삼성전자우	74400.0	상승 4,400	+6.29%	100.0	6976077.0	516986.0	5751012.0	74200.0	74900.0	73500.0	74300.0	74400.0
5	한화에어로스페이스	1042000.0	하락 55,000	-5.01%	5000.0	288611.0	302069.0	189175.0	1069000.0	1079000.0	1027000.0	1041000.0	1042000.0
6	두산에너빌리티	74500.0	상승 9,700	+14.97%	5000.0	19302308.0	1374733.0	6999933.0	66300.0	74800.0	66100.0	74500.0	74600.0
7	HD현대중공업	515000.0	하락 13,000	-2.46%	5000.0	213684.0	110444.0	196437.0	522000.0	524000.0	509000.0	515000.0	516000.0
8	현대차	217000.0	하락 3,000	-1.36%	5000.0	779161.0	169536.0	1099674.0	219000.0	220250.0	216500.0	216500.0	217000.0
9	KB금융	112800.0	하락 4,000	-3.42%	5000.0	1692490.0	191704.0	1308296.0	113500.0	115000.0	111500.0	112700.0	112800.0

- **number of stocks**
  - 4,192
- **Total columns**
  - 33
- **Data composition**
  - Market : 15
  - Financial : 17
  - Category : 1

- The project focuses on the domestic capital market, primarily the **KOSPI** and **KOSDAQ** indices
  - After building the core system, the framework is designed to be scalable to **other markets**, such as the **S&P 500**

ROADMAP


너무 미래 신용평가
안 좋은 기업이 있으면
return
Company Trend
User Request
Generate Portpolio
credit rating
Output
Portpolio & Report
Company
Field Trend
Market
Data
News
Data
Financial
Data

Market Data

STOCK DATA

BeautifulSoup - Naver Crawling
FinanceDataReader

FINANCIAL DATA

FinanceDataReader
ts2000

MERGE

PREPROCESSING

Data clensing
Grouping and Removing
Encoding and transformation
Missing Value - handling

FEATURE ENGINEERING

RSI
Bollinger_band
Moving Average
ROE / PER / PBR
.
.
.

News Data

각 기업명을 키워드로 매일 뉴스 수집 → **중복 제거 및 MYSQL 저장**

- 기업명
- 기사 ID
- 기사 게재일
- 기사 제목
- 기사 내용

Sentence_BERT 모델의 의미 유사도 분석을 통한 추출적 요약 →

NEO4J 노드 저장

- 기업명
- 기사 ID
- 요약된 기사 내용

기업 별 대표 문장 생성

추출적 요약된 문장의 집합을 다시 요약하여 각 기업의 데이터로 활용

코사인 유사도

전체 뉴스 수집 → **중복 제거 및 MYSQL 저장**

- 기사 ID
- 기사 게재일
- 기사 제목
- 기사 내용

Sentence_BERT 모델의 의미 유사도 분석을 통한 추출적 요약 →

NEO4J 노드 저장

- 기업명
- 기사 ID
- 요약된 기사 내용

각 기업과 링크 연결

추출적 요약된 문장의 집합을 다시 요약하여 각 기업의 데이터로 활용

Relational Database (MySQL)

COMPANY_TABLE

Company_ID

Company_NAME

COMPANY_FINANCE

Company_ID

Year_Quater

financial informations

COMPANY_STOCK

Company_ID

주식시장 내용

ARTICLE_TABLE

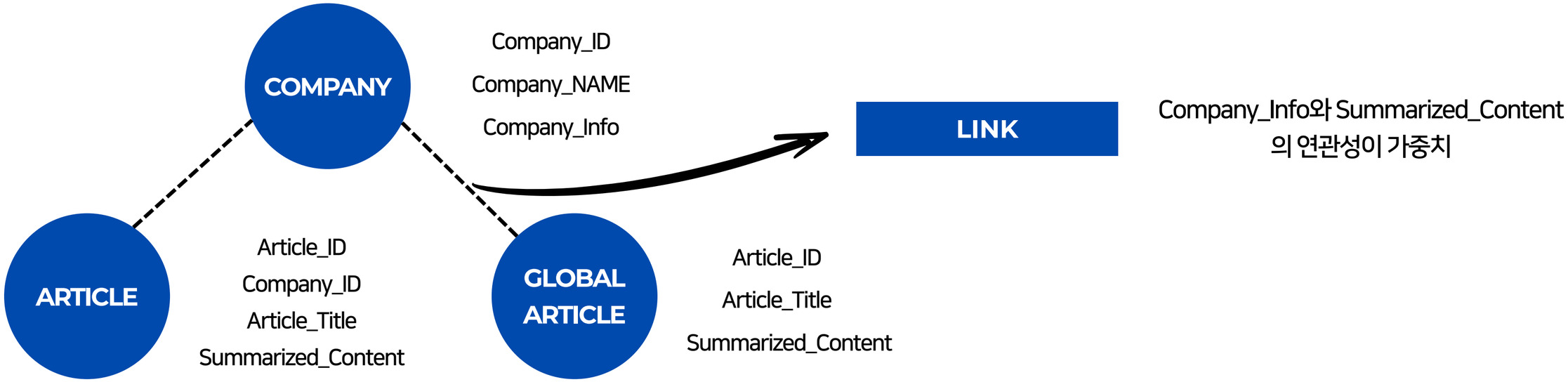
Article_ID

Posted_Date

Article_Title

Article_Content

Graph Database (Neo4j)



Future Works

Integrated Analysis and Summary Text Generation

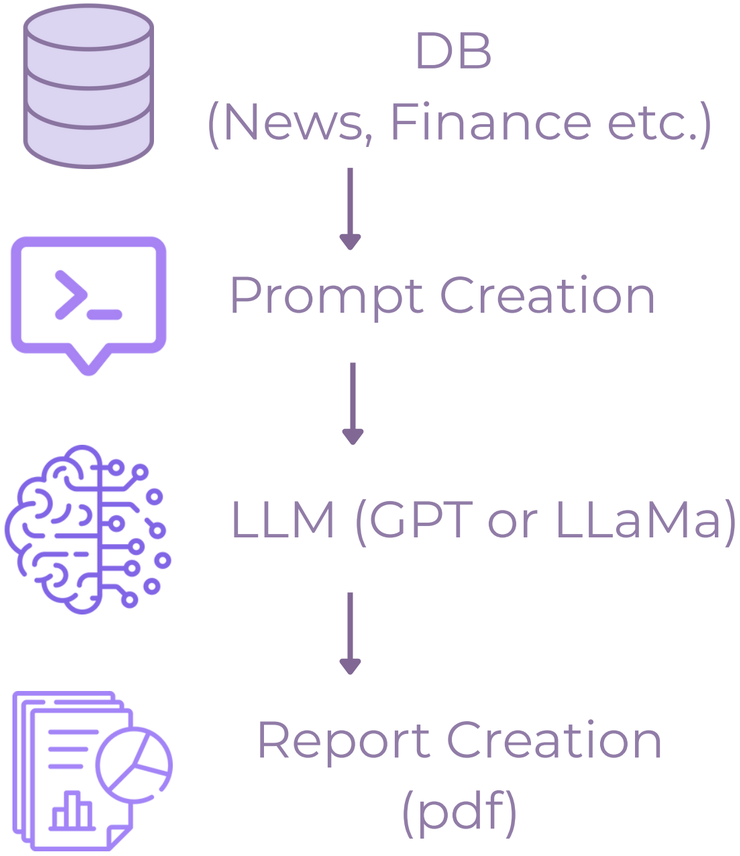
- LLM provides a comprehensive interpretation of individual analysis results to describe the overall trends and insights of the company in natural language

Implementation of LLM Input Prompt Generator

- Develops a function that automatically structures data extracted from the database and converts it into a standardized prompt that LLM can understand

Analytical data-based report generation module

- Write script code that generates automatic reports based on integrated analysis results



EXPECTED OUTCOMES

Portfolio

Recommended Portfolio

1. 삼성전자
   추천비중 - 20%
2. 하이닉스
   추천비중 - 20%
3. 메리츠증권
   추천비중 - 15%

Reports

삼성전자

1. 뉴스 Trend
   - 영업익 10조, 10만전자 달성 촉각...
2. 관련분야 (sector, 경쟁업체 trend)
   - 중국 수출 통제 강화...
3. 재무 데이터
   - N년치 재무제표 데이터
4. 감성분석 데이터
   - 뉴스 기반 감성분석
   - 영업익 10조 달성 → 긍정