LinguaTech Expo with

Multimodal Stock Analysis

팀원

헤더 : 이현우

팀원 : 강현규, 민건우

AIEYES

AI 융합대학 인공지능 학회 AIEYES. AI팀

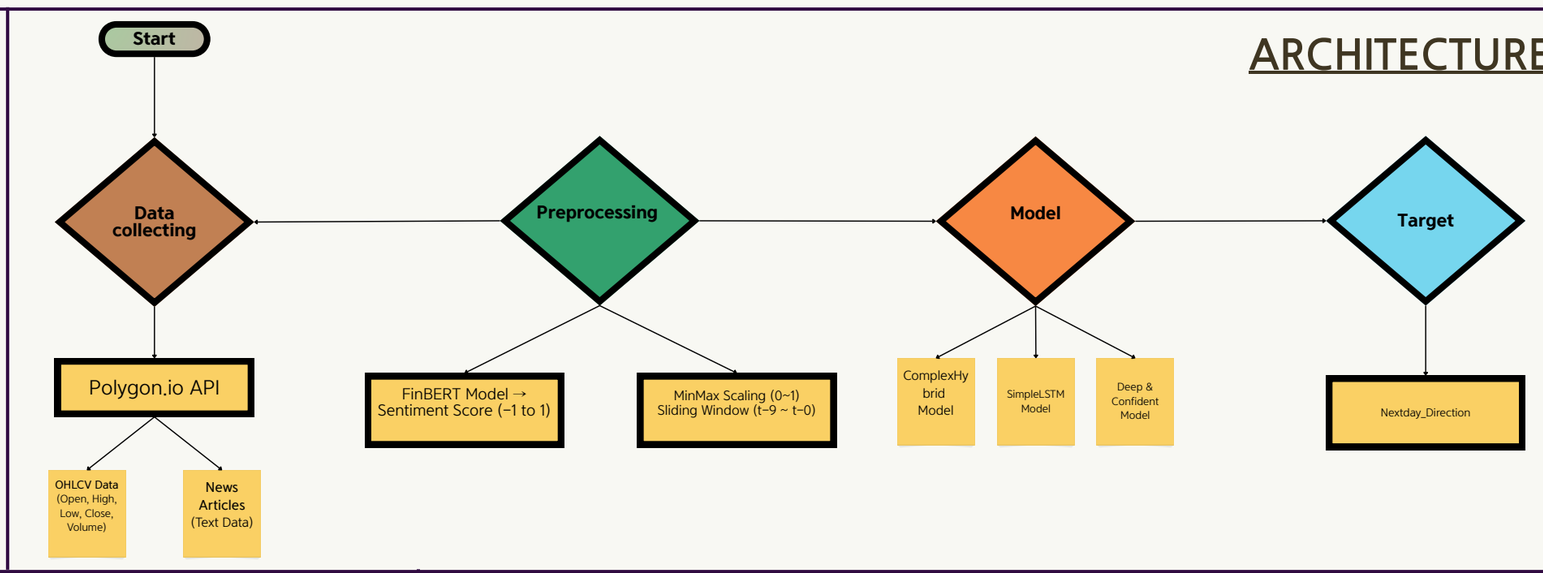
BACKGROUND

금융 시장 데이터를 활용한 딥러닝 기반 주가 예측 모델 개발은 학계와 산업계 모두에서 오랫동안 추구해 온 과제입니다. 그러나 본 연구팀은 다양한 아키텍처와 손실 함수를 실험하는 과정에서 "완전한 주가 예측의 본질적 불가능성"을 직면하게 되었습니다.
초기 단계에서 우리는 BiLSTM, GRU, Multi-Head Attention을 결합한 하이브리드 모델부터 시작하여 단순 LSTM, 가중치 기반 손실 함수 등 다양한 접근을 시도했습니다
그러나 복잡한 모델은 제한된 학습 데이터(338일)에서 과적합(Overfitting)을 피할 수 없었고, 단순한 모델과 표준 MSE 손실 함수는 모델이 평균값으로만 회귀하는 "겁쟁이 모델"을 만들어냈습니다.
이는 주가 예측에서 흔히 나타나는 현상으로, 모델이 Loss를 최소화하기 위해 "내일 가격은 오늘 가격과 같다"는 무의미한 결론을 내리는 것입니다.
이러한 반복된 실패 경험은 중요한 인사이트를 제공했습니다.
주가의 정확한 가격 수준을 예측하는 것은 시장의 높은 노이즈 비율(95% 이상)과 무작위성(Random Walk) 때문에 근본적으로 제한적이라는 것입니다.
따라서 우리는 연구 방향을 전환하여, 가격의 절댓값 예측 대신 "고변동성 구간의 방향성 포착"에 집중하기로 결정했습니다.
이는 실전 거래에서 모든 시점의 예측보다 "확실한 신호가 있는 순간"만을 정확히 포착하는 것이 더 실용적이라는 판단에서 비롯되었습니다.

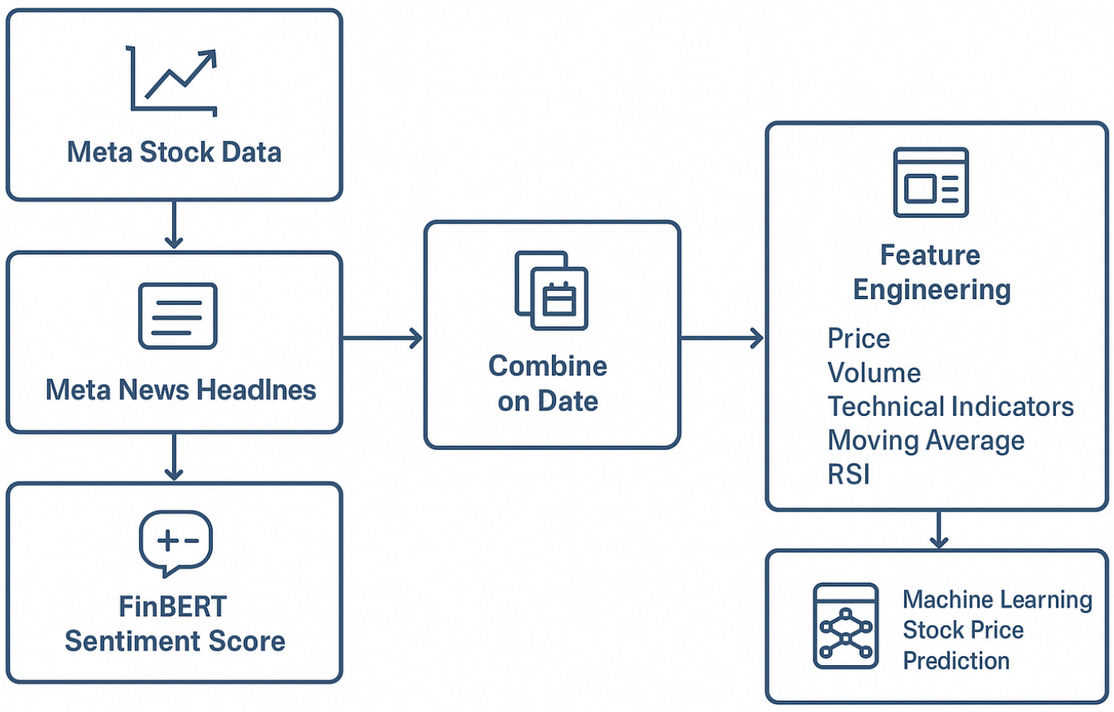
PURPOSE

금융 시장 데이터에 대한 딥러닝 기반 예측 모델의 개발은 높은 비선형성과 노이즈로 인한 근본적인 도전 과제를 제시합니다. 본 연구는 Bi-LSTM, GRU, Attention 메커니즘을 포함한 고도화된 하이브리드 아키텍처로부터 출발하였으나, 초기 가설과는 달리 모델의 복잡도 증가가 일반화 성능의 저하를 초래함을 발견하였습니다. 특히 과적합(Overfitting), 평균값 회귀(Mode Collapse), 맹목적 과신(Overshooting)이라는 핵심 실패 사례를 식별하고 이를 극복하기 위해 (1) 피처 엔지니어링 최적화, (2) 집중 손실 함수(Focus Loss), (3) 엔트로피 페널티를 통한 예측 결과를 제안합니다. 본 방법론을 통해 모델이 “모든 구간의 예측"에서 벗어나 "확실한 신호의 포착"에 집중하는 것이 목표입니다.

ARCHITECTURE



데이터 수집 및 전처리



데이터 수집 및 전처리

- Financedatareader 라이브러리로 주가 데이터를 그리고 beautiful soup으로 메타 관련 뉴스 기사를 수집.
- 5년간의 일별 주가(OHLCV)와 회사·언론 뉴스로부터 수집한 헤드라인을 날짜 기준으로 결합.
- 뉴스 텍스트는 금융 특화 언어모델(ProsusAI/finbert)을 통해 -1~+1 범위의 감정 점수(sentiment_score)로 변환.
- 가격·거래량·기술지표(MA, RSI)와 감정의 이동평균·모멘텀·변동성·지연값을 추가로 계산하여 총 34개의 피처를 구성.
- 감정과 가격·거래량의 상호작용 항을 포함해 투자 심리와 가격 움직임 간 비선형 관계를 포착하도록 설계.
- 최종적으로 (X: 23개 피처, y: 다음날 수익률 및 상승 여부) 형태의 CSV 데이터셋을 구성하여 머신러닝 기반 주가 예측 모델의 입력으로 사용.

파생변수 제작

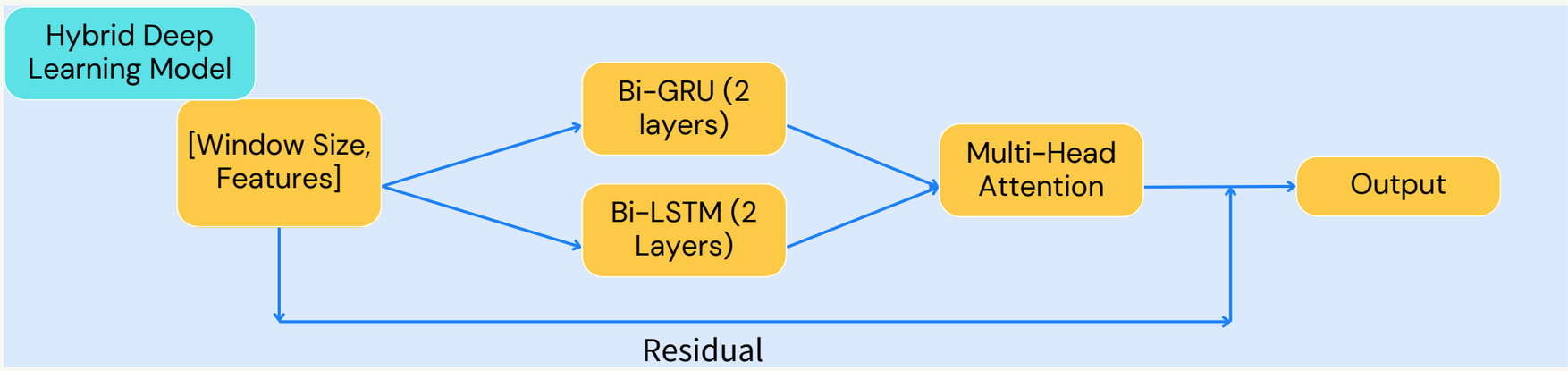
1. 감정(Sentiment) 피처
sentiment_score: 뉴스 감정 점수 (-1~+1) [Tetlock 2007]
sentiment_ma_3/7/14: 감정 이동평균 (단/중기 트렌드) [Murphy 1999]
sentiment_momentum: 감정 변화율 [Jegadeesh 1993]
sentiment_lag_1/3/7: 감정 지연값 (정보 반영 지연) [Fama 1970]

2. 거래량(Volume) 피처
volume_ma_5: 5일 평균 거래량 (기준치)
volume_ratio: 거래량 비율 (정상=1.0) [Karpoff 1987]
volume_change_pct: 거래량 변화율 (참여 강도)

3. 기술지표(Technical) 피처

ma_5/10/20/50: 가격 이동평균 (추세) [Murphy 1999]
rsi_14: 상대강도지수 (과매수/과매도) [Wilder 1978]
volatility: 일중 가격 변동성 [Parkinson 1980]

4. 타겟 피쳐 : next_day_return (다음 날 수익률)
df['next_day_return'] = df['close'].pct_change().shift(-1)



First model

총 레이어 수: 30
총 파라미터 수: 333,841

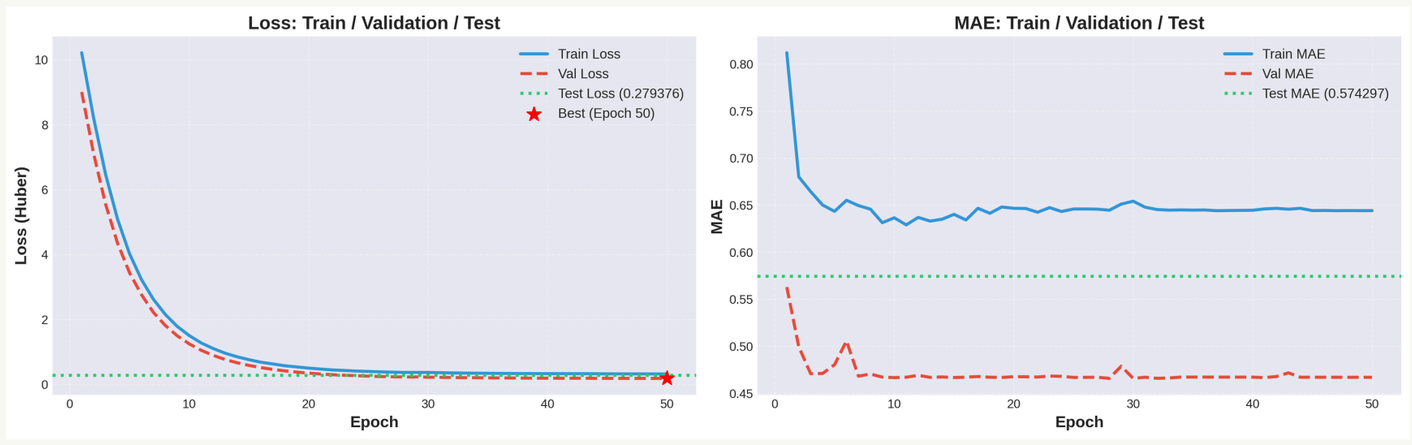
Branch 1: 장기 의존성 학습 (Bi-LSTM)
bidirectional=True: 데이터를 과거에서 미래로, 미래에서 과거로 양방향으로 처리하여 문맥 정보를 더 풍부하게 학습.

Branch 2: 단기 변동성 학습 (Bi-GRU) lSTM보다 구조가 간단하여 연산 속도가 빠르고 단기적인 변동성에 민감하게 반응. > 갑작스러운 가격 변화나 노이즈와 같은 단기적인 특징을 포착.

Feature Fusion & Attention (특징 결합 및 집중) : LSTM과 GRU가 추출한 서로 다른 종류의 정보가 결합
MultiheadAttention: 결합된 특징 맵에서 중요한 부분(Key Events)에 가중치를 부여.

Residual Connection & Global Pooling (잔차 연결 및 풀링) : Attention을 통과한 정보와 원래 정보를 더해줌(ResNet 구조). 이는 정보 손실을 막고 학습을 안정화하는 역할.
Prediction Head (최종 예측) : 배치 정규화와 드롭아웃을 사용하여 과적합을 방지하고 학습 성능을 높임. 최종적으로 다음 날 등락률을 출력

Result & Failure



학습 정체: Epoch 10 이후 개선 없음
→ 실제 최적 모델은 Epoch 10~15 부근

절대 성능: MAE 0.574는 중간 수준, RMSE : 0.53 으로 평균 성능

Train MAE가 높음: 학습 데이터 난이도 문제
→ 데이터 품질 재검토
→ 노이즈 제거 또는 이상치 처리

LOGIC

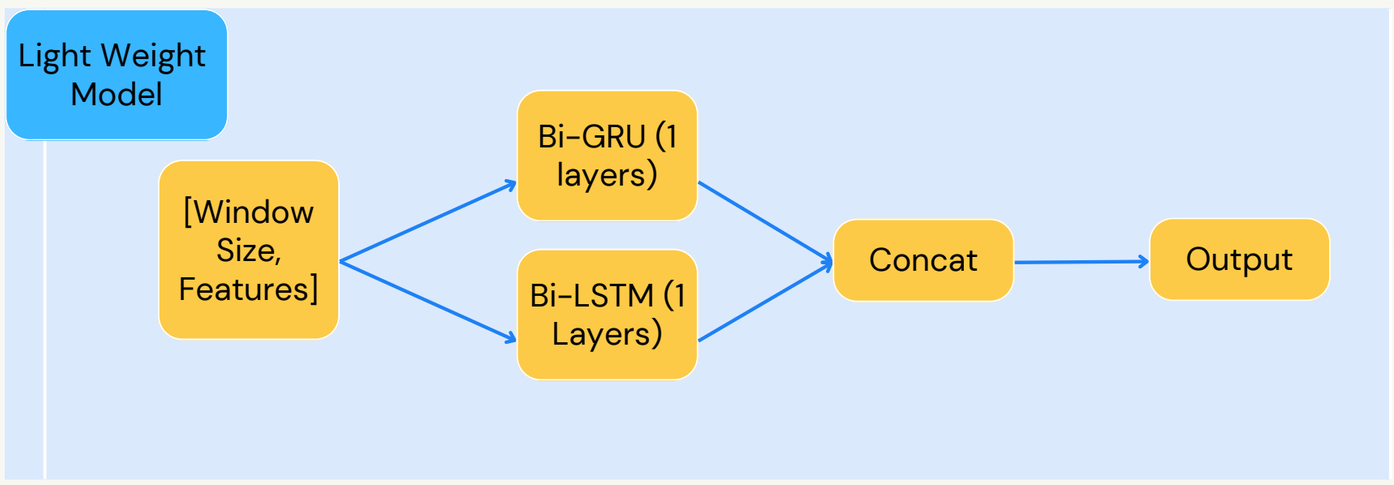
1. 듀얼 브랜치 (LSTM Branch + GRU Branch)
- 병렬 처리로 장기+단기 변동성 및 다양한 패턴 학습

2. Multi-Head Self-Attention
- 시계열의 중요한 시점 자동 감지
- 장기 의존성 강화

3. Residual Connections (Skip Connections)
- 그래디언트 소실 방지
- 깊은 네트워크 안정화

4. Layer Normalization
- 학습 안정성 향상
- 빠른 수렴

5. Ensemble Dense Layers
- 비선형 변환 강화
- 과적합 방지



Second model

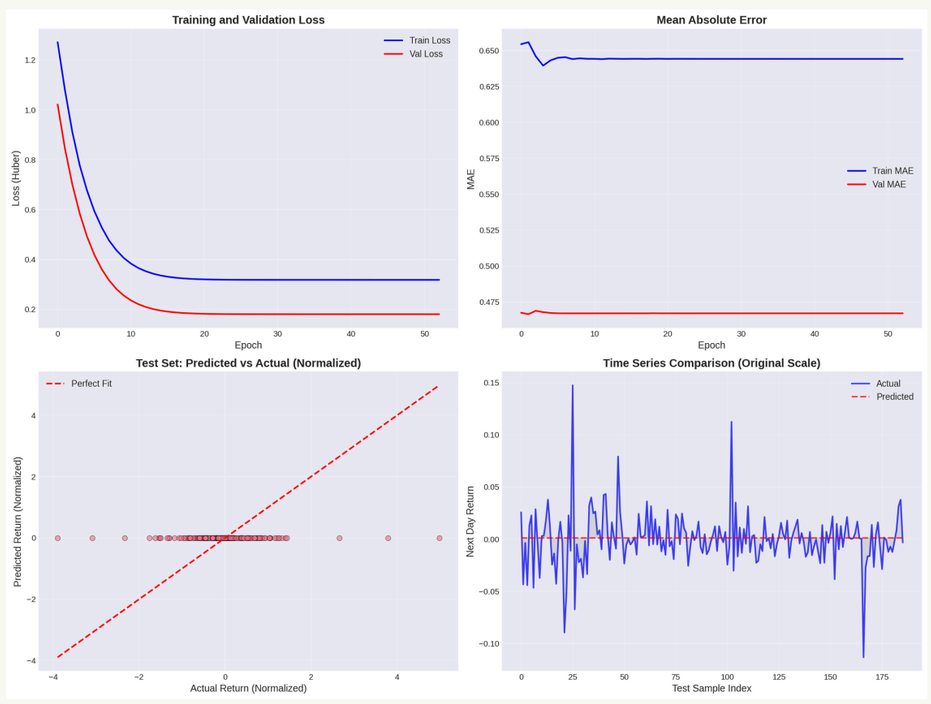
1. 이전의 고급 모델에 비해 경량화 모델은 총 파라미터 수를 40만 개에서 4만 개로 1/10 수준으로 대폭 축소.

2. 이전 모델이 LSTM과 GRU 각각 2층 구조(64→32 유닛, 48→24 유닛)였던 것과 달리, 경량화 모델은 각각 단일층(LSTM 32 유닛, GRU 24 유닛)으로 단순화하여 계층적 복잡도를 제거하고 과적합 위험을 낮춤.

3. 이전 모델에서 시계열의 중요 시점을 자동 감지하던 Multi-Head Attention(4개 헤드)과 그래디언트 소실 방지를 위한 Residual Connection, 학습 안정화를 위한 8개의 Layer Normalization을 모두 제거하여, 소규모 데이터셋에서 노이즈를 학습하는 문제를 원천 차단.

4. 이전 모델의 Dropout 비율 0.1과 L2 정규화 0.01에 비해, 경량화 모델은 Dropout을 0.4~0.5로 4~5배 강화하고 L2 정규화를 0.02로 2배 증가시켜, 단순화된 구조에서도 강력한 과적합 방지 메커니즘을 확보.

Result & failure



1. Training and Validation Loss
문제점:
Train Loss: 0.32까지 하강 (양호)
Val Loss: 0.18에서 정체 (10 epoch 이후)
Gap이 약 2배: 전형적인 과적합 신호
원인: 모델이 훈련 데이터의 노이즈까지 학습하고 있으나 검증 데이터에는 일반화하지 못함

2. Mean Absolute Error
Train MAE: 0.645로 시작 후 전혀 개선 안 됨 (거의 수평선)
Val MAE: 0.465로 시작 후 역시 정체
의미: Huber loss는 감소하지만 실제 예측 정확도(MAE)는 개선되지 않는 현상 - 모델이 이상치 처리에만 집중하고 실제 패턴 학습은 실패

3. Predicted vs Actual
모델 붕괴 현상:
모든 예측값이 y=0 근처에 밀집 (빨간 점들이 가로선 형태)
실제값은 -4 ~ +6 범위인데 예측은 거의 0
모델이 항상 평균값(0)만 예측하는 "평균 회귀" 상태
원인: 과도한 regularization(L2=0.02) + 높은 Dropout(0.4-0.5)로 모델이 보수적으로 수렴

4. Time Series Comparison
파란선(실제): 큰 변동성 (-0.10 ~ +0.15)
빨간선(예측): 거의 0에 고정 (변동 없음)
모델이 "항상 0%의 변동성 "를 예측하는 무용지물 상태

LOGIC

초기 다음날 주가 등락률(연속값) 예측 > 상승/하락 방향성 예측(이진 분류) 목표 전환.
이를 해결하기 위해 Weighted Loss (Beast Mode) 전략 도입
1. 예측값의 절대값이 실제값보다 작을 경우 손실 가중치를 30배 증폭시켜 모델이 과감한 예측을 수행하도록 강제.
2. Weight Decay를 0으로 설정하여 모든 규제를 제거함으로써 파라미터 업데이트의 제약을 완전히 해제.

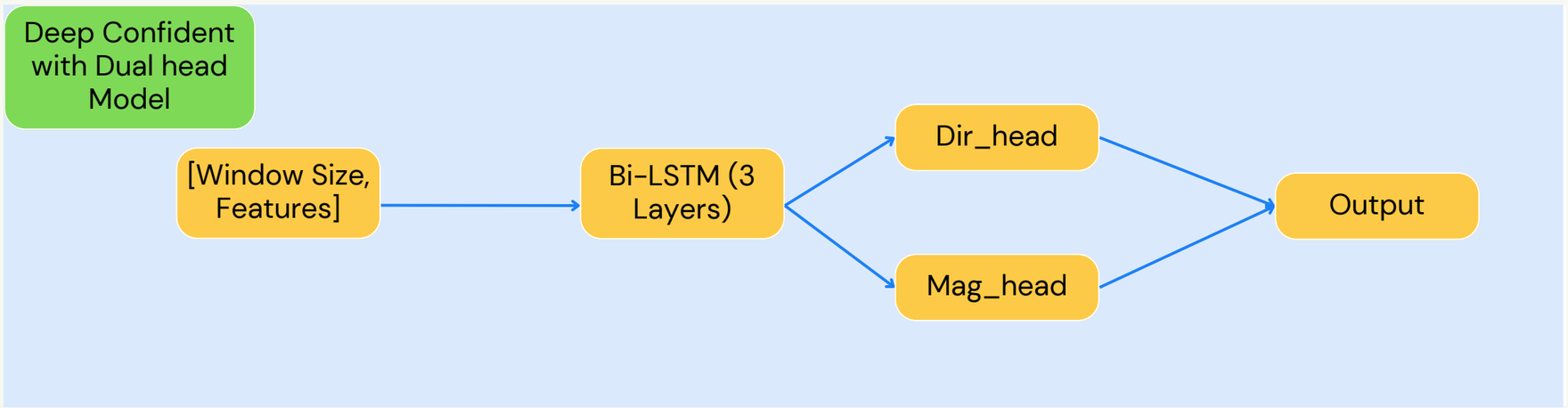
```
undershoot_mask = (torch.abs(pred) < torch.abs(y_b)).float()

weights = 1.0 + (torch.abs(y_b) * 30.0 * undershoot_mask)

mse_loss = (pred - y_b) ** 2
loss = (mse_loss * weights).mean()


Gusrb10001
```

LOGIC RESULT

Pros: 모델이 예측을 평균값만 하는 평균값 회귀 현상을 어느 정도 극복하고
모델 예측값의 변동성이 커짐

Cons: 변동성 크기만 증폭시키고 방향 판단 능력은 마비시켜, 방향 정확도가 30%대로 추락

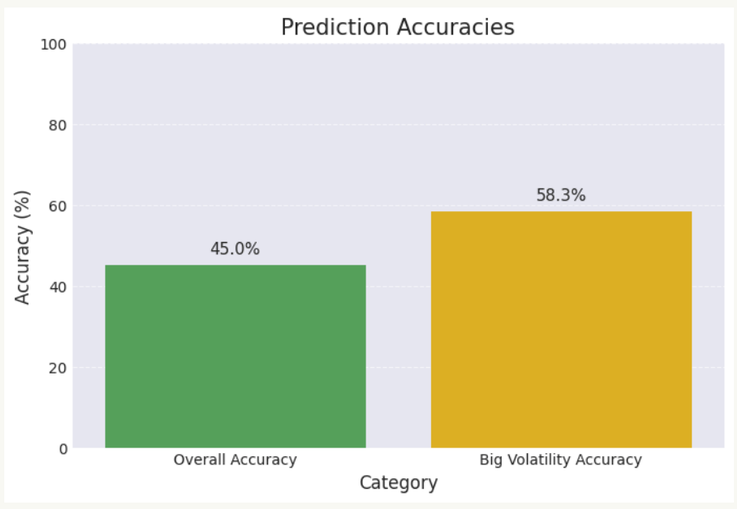
THIRD MODEL



DeepConfidentModel은 주가 방향성과 변동 크기를 동시에 예측하는 Multi-task Learning 구조로 설계. 모델의 백본(backbone)은 3-layer Bidirectional LSTM으로 구성되어 있으며, 양방향 구조를 통해 과거와 미래의 시계열 패턴을 모두 포착. 각 LSTM 층은 64개의 hidden units을 가지며, 양방향 처리로 인해 최종 출력은 128차원의 feature representation.
LSTM의 마지막 timestep 출력은 BatchNorm1d 층을 통과하여 학습 안정성을 확보하며, 이후 Dropout(0.5)을 적용하여 과적합을 방지.

Dual-Head 구조
정규화된 feature는 두 개의 독립적인 prediction head로 분기. 방향 Head는 2개의 fully-connected layer(128→64→1)와 GELU activation으로 구성되어 있으며, 로짓(logit) 값을 출력하여 상승/하락 이진 분류를 수행. 크기 Head는 동일한 구조에 Softplus activation을 추가하여 항상 양수인 변동폭을 회귀 예측. Softplus 함수는 물리적으로 음수가 불가능한 수익률 절대값을 보장하며, 부드러운 미분 특성으로 안정적인 학습 .

RESULT & FAILURE



전체 예측률은 낮지만 큰 변동성을 약 60%의 정확도로 포착

LOGIC

Confident Loss Function

방향성과 변동 크기를 동시에 학습하면서 확신도를 강제하는 복합 손실 함수. 손실 함수는 세 가지 핵심 컴포넌트로 구성.

1) 변동성 기반 동적 가중치

실제 수익률의 절대값에 비례하는 가중치를 계산하며, 최소값 0.5에서 clamp되어 3배로 증폭 > 변동성이 큰 시점(예: 주요 뉴스 발표일)의 예측 오차에 더 큰 페널티를 부여하여 모델이 중요한 시장 전환점에 집중하도록 유도.

2) 방향 손실 (Weighted BCE)

Binary Cross-Entropy with Logits Loss를 사용, 앞서 계산된 가중치를 각 샘플에 개별적으로 적용(reduction='none') > 상승/하락 이진 분류의 정확도를 직접적으로 최적화하며, 가중치 적용으로 평균적인 시점보다 변동성이 큰 시점의 방향 예측 정확도를 우선시.

3) 엔트로피 페널티

Sigmoid 확률값으로부터 Shannon entropy를 계산하고 0.5배의 계수로 손실에 추가. 엔트로피는 -(p*log(p) + (1-p)*log(1-p)) 형태로 계산되며, 확률이 0.5 근처(애매한 예측)일 때 최대값을 가지므로 이를 페널티로 부여하여 모델이 0 또는 1에 가까운 확신 있는 예측을 하도록 강제. > (확신이 없다면 예측하지 말라).

4) 크기 loss (MSE)

예측한 변동폭과 실제 변동폭의 절대값 간 평균 제곱 오차를 계산. Softplus 활성화로 인해 항상 양수인 pred_mag와 target_mag 사이의 회귀 손실을 최소화하여 수익률의 규모를 정확히 예측하도록 학습.

최종 loss는 세 컴포넌트의 단순 합산(dir_loss + entropy_loss + mag_loss)으로 계산되며, 각 요소가 균등한 중요도로 기여.

학습 루프 최적화

1) Gradient Clipping

역전파 과정에서 gradient norm이 1.0을 초과하면 자동으로 스케일링하여 학습 불안정성을 방지. 금융 데이터의 급격한 변동성으로 인한 gradient explosion을 효과적으로 억제.

2) 동적 Learning Rate 조정

ReduceLROnPlateau scheduler는 매 epoch마다 train loss를 모니터링하며, 개선이 멈추면 learning rate를 절반으로 감소시켜 local minima 주변에서 세밀한 최적화를 수행.

3) 평가 메트릭: 스마트 스케일링 및 큰 파도 적중률

평가 단계에서는 확신도 기반 적응적 스케일링을 적용. 예측 확률에서 0.5를 뺀 절대값에 2를 곱해 0~1 범위의 confidence를 계산하고, 이를 제곱하여 확신도가 높을 때만 예측 크기를 증폭시킴. 또한 실제 수익률이 1.5% 이상인 "큰 파도" 구간을 별도로 추적하여, 실질적 수익 기회가 있는 구간에서의 방향 적중률을 독립적으로 평가.

DATA UTILIZATION VERIFICATION

배경 및 문제 정의
뉴스 텍스트의 감성 정보(FinBERT)와 시장 데이터(Price, Volume)를 결합한 LSTM 기반 주가 예측 모델 구축. 초기 모델링 단계에서 Loss 정체 및 과소적합 현상이 지속적으로 관측됨.

도입 목적
딥러닝(LSTM)의 블랙박스 문제를 배제하고, 모델의 구조적 결함이 아닌 데이터 자체의 유효성을 검증하기 위함.
데이터 내에 학습 가능한 신호가 존재하는지, 단순 노이즈인지를 판별하는 '데이터 디버깅' 도구로 활용.

VERIFICATION TOOLS & INTERPRETATION METRICS

검증 모델: XGBoost (Extreme Gradient Boosting)

- 정의 (Definition)
  - 결정 트리 기반의 앙상블 알고리즘으로, 이전 트리의 오차를 순차적으로 보정하며 학습하는 모델.
- 선정 이유
  - 정형 데이터 분석의 표준 모델로, 과적합 방지와 특성 중요도(Feature Importance) 파악에 탁월하여 데이터 검증에 최적화됨.
  - 내부 연산을 알 수 없는 딥러닝 대신, 설명 가능한(Explainable) 모델을 Baseline으로 도입하여 분석의 투명성 확보.

해석 지표의 이원화: Feature Importance vs. SHAP

- 데이터의 결함을 다각도로 포착하기 위해 두 가지 해석 기법을 비교 분석.
  - Feature Importance
    - 기준: 정보 이득 (Information Gain).
    - 의미: 모델이 학습 과정에서 정보를 얻기 위해 특정 데이터를 '얼마나 자주 사용했는가'를 나타냄.
    - 한계: 변수의 중요도(크기)는 알 수 있으나, 결과에 미친 방향성(긍정/부정)은 알 수 없음.
  - SHAP Value
    - 기준: 섀플리 값 (Game Theory 기반).
    - 의미: 특정 변수가 실제 예측값(수익률)을 '양(+) 혹은 음(-)으로 얼마나 변화시켰는가'를 나타냄.
    - 장점: 변수의 실질적인 영향력과 인과관계의 구체적인 방향성을 동시에 파악 가능.

ANALYSIS

- 지표 간의 괴리 해석
  - feature_importance 와 shap의 괴리

모델은 sentiment_momentum(뉴스 심리)을 가장 중요한 변수로 선택했으나, SHAP 분석 결과 해당 변수의 영향력은 방향성이 혼재(Mixed)되어 나타남.
즉, 뉴스 데이터가 실제 주가 방향성을 설명하지 못하고 있음에도 모델이 허위 양성(False Positive)으로 학습함

  - 시계열 인과관계의 붕괴

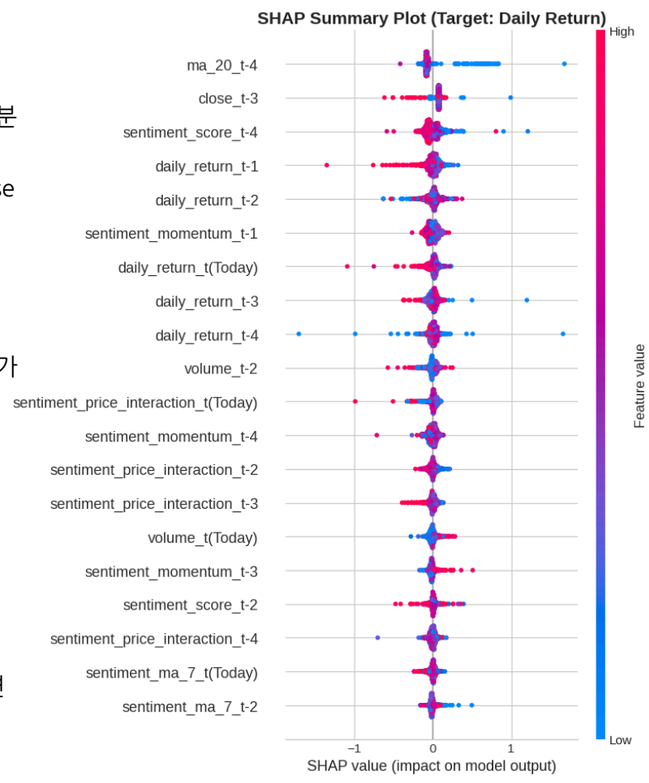
SHAP 분석에서 가장 영향력이 큰 변수로 직전일(t-1)이 아닌 4일 전 데이터(ma_20_t-4)가 도출됨
금융 시계열에서 4일 전 정보가 내일 주가에 가장 큰 영향을 미치는 것은 논리적으로 불가능하며, 이는 모델이 우연한 노이즈에 과적합되었음을 의미함

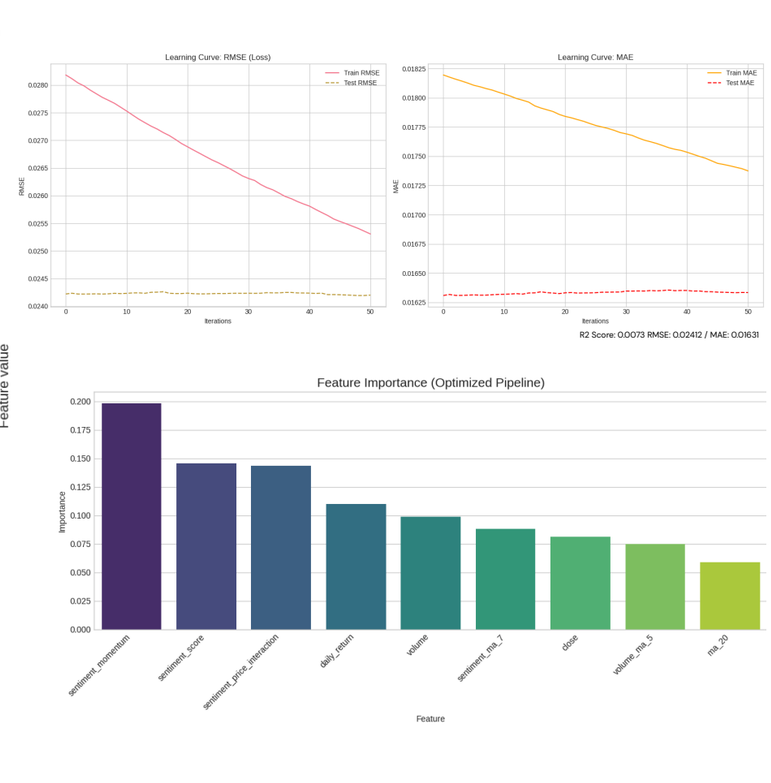
- 결론
  - 진단 결과 및 향후 계획

최종 진단: 데이터 무결성 결함
학습 실패의 원인은 알고리즘 문제가 아닌 데이터셋의 구조적 결함에 있음.
특히 t-4 변수 의존 현상은 데이터 전처리 과정에서의 치명적인 시점 불일치를 시사함

  - 향후 계획

모델 튜닝 중단: 데이터 엔지니어링 단계로 회귀.
Time-Lag 전수 조사: 뉴스 데이터와 주가 데이터 병합(Merge) 시 날짜 매핑 및 Shift 연산 오류 육안 검증.
타겟 재정의: 랜덤 워크 성향을 줄이기 위해 예측 타겟을 수치 예측에서 '방향성(Up/Down)' 분류로 단순화 고려





CONCLUSION

저희 AI 팀은 딥러닝 기반 주가 예측 모델 개발 과정에서 완전한 주가 예측의 본질적 한계를 직면하게 되었습니다. BiLSTM, GRU, Multi-Head Attention 하이브리드 모델부터 단순 LSTM_GRU, 가중치 기반 손실 함수까지 다양한 아키텍처를 실험했으나, 복잡한 모델은 제한된 데이터(338일)에서 과적합되고 단순 모델은 평균값으로만 회귀하는 "겁쟁이 모델"이 되었습니다. XGBoost Feature Importance 분석 결과, 모델이 t-4(4일 전) 변수에 과도하게 의존하는 현상이 발견되어 알고리즘이 아닌 데이터셋의 구조적 결함이 핵심 원인임을 확인했습니다. 뉴스 감성 데이터와 주가 데이터 병합 과정에서의 시점 불일치가 치명적인 데이터 누수를 초래했을 가능성이 높다고 판단됩니다.
주식 시장은 본질적으로 95% 이상의 노이즈와 Random Walk 특성을 가지며, 정치적 사건, 자연재해, 기관 투자자 내부 거래 등 예측 불가능한 외부 요인에 크게 영향받아 가격의 절댓값 예측은 근본적으로 제한적입니다.
따라서 연구 방향을 "고변동성 구간의 방향성 포착"으로 전환하여, 모든 시점보다 "확실한 신호가 있는 순간"만 정확히 포착하는 실용적 접근을 채택했습니다만 결과는 그리 좋지 못했습니다. 따라서 본 연구는 과적합, 평균값 회귀, 맹목적 과신이라는 세 가지 핵심 실패 패턴을 식별하고, 피처 엔지니어링 최적화, 집중 손실 함수, 엔트로피 페널티를 결합한 학습 프레임워크를 시도해보았다는 것에 의의를 두고 있습니다. 저희는 "불확실성이 높은 금융 도메인에서는 완전한 예측보다 신뢰할 수 있는 부분 신호를 포착하는 것이 더 가치 있다"라고 생각합니다. 우리의 실패 경험들은 AI 금융 예측의 현실적 한계와 실용적 대안을 제시하는 솔직한 인사이트로 전환되었습니다.

NEXT STEP

1.모델 튜닝 및 파라미터 조정 중단
복잡한 딥러닝 모델의 추가적인 튜닝은 학습 데이터가 충분치 않을 때 실질적인 개선을 기대하기 어렵기 때문에 잠정 중단합니다
.

2. 데이터 엔지니어링 및 검증 단계로 전환
현재 데이터셋의 시점 불일치 문제(time-lag mismatch)와 데이터 누수(data leakage)가 가장 큰 장애요소로 진단되었으므로, 데이터 구조와 전처리 과정을 전수 조사합니다.

3. 외부 변수·Global 데이터 병합 테스트환율, 금리, 이벤트 등 글로벌 외생 변수와 기존 데이터의 조합 실험을 추가하여, 단기 움직임뿐만 아니라 전체 시장 환경에 대응할 수 있는 구조를 모색합니다.

4.데이터 품질 및 정확도 평가 프로세스 구축
모든 데이터 파이프라인에 대해 정확도·무결성 점검 프로세스(샘플링, 실제 변동률 비교 등)를 구축합니다.

5. 소규모 트레이딩 시뮬레이션으로 실전 검증
전처리 및 데이터 품질 개선 후, 소규모 실전 모의 매매(백테스트)를 진행하여 분석 결과의 실효성을 점검합니다.