



CAPSTONE DESIGN

Knowledge Distillation 기반

금융 도메인 특화

감성 분류 AI 모델 경량화

21조 | 김진희 이가연 이예진 이현수 조수민

문제 정의 및 목표

Teacher-Student 구조와 Knowledge Distillation을 활용한 도메인 특화 경량화 AI 모델 개발

거대 모델 (Teacher)

고성능 — 금융 감성 자동화 가능

금융 기사 분석 성능은 높지만 실시간 서비스 적용이
어려움

GPU 의존도와 추론 비용이 높아 운영 부담 발생

소형 모델 (Student)

가벼움 — CPU 실시간 동작 가능

단독 학습 시 금융 도메인 맥락 이해가 부족함
투자 판단에 중요한 미세한 감성 신호 포착 한계

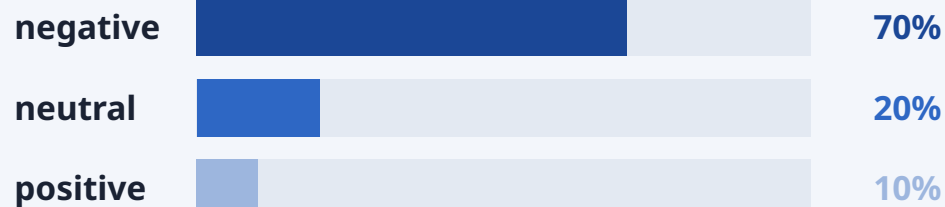
GOAL

고성능 금융 감성 분석 모델의 지식을 경량 모델에 전이하여,
CPU 환경에서도 실시간으로 동작 가능한 금융 뉴스 분석 시스템 구현

Knowledge Distillation 방법론

일반 fine-tuning → 정답 하나만 학습 / KD → Teacher가 부여한 **클래스별 확률 분포(Soft Label)**까지 함께 학습

예시 · Teacher의 판단 (Soft Label)



CE Loss

CrossEntropy

정답 라벨
Hard Label

+

KD Loss

KL Divergence

Teacher 분포
Soft Label

$$\text{Total Loss} = \alpha \cdot \text{CE} + (1-\alpha) \cdot \text{KD}$$

$\alpha = 0.8$

정답 80%

Temperature = 1.5

분포 평활

Best checkpoint

val Macro F1

Stratified split

분포 보존

전체 실험 프로세스

01

Teacher 후보 선정

Claude Sonnet vs. FinBERT

02

영어 KD 검증

FinBERT → DistilBERT / MiniLM

03

한국어 KD 적용

KorFin-ASC 재구축 KR-FinBERT → KoELECTRA

04

KD 성공 검증

Teacher-Student 일치도

05

추가 경량화

Pruning + max_len 최적화

06

Chrome 확장 프로그램 구현

네이버 뉴스 실시간 분석

핵심 흐름: 데이터·모델이 풍부한 영어에서 KD 작동 검증 → 데이터가 부족한 한국어 금융 도메인에 안전하게 이식

STEP 1

Teacher 후보 선정

Model	Params	Macro F1	비고
Claude Sonnet	~175B+	0.8978	범용 LLM
FinBERT	110M	0.9365	영어 / 금융 특화

도메인 특화 판단이 중요한 태스크 기준에서는
범용 LLM보다 **도메인 학습 모델**이 더 Teacher 모델에 적합하다고 판단



동일 논리로 한국어 Teacher 모델을 **KR-FinBERT (110M, F1 0.803)** 선정

선정 결과 · Teacher

FinBERT

0.9365

Macro F1

도메인 특화 > 범용 LLM

STEP 2

영어 환경에서 KD 구조 검증

DATASET

Financial PhraseBank
75% agree / 3,453개

클래스 불균형
neutral 62% → class weight 적용

지표 = Macro F1
전부 neutral 예측해도 Acc 62%
→ 클래스 균등 평균 F1이 신뢰성↑

RESULT · Macro F1

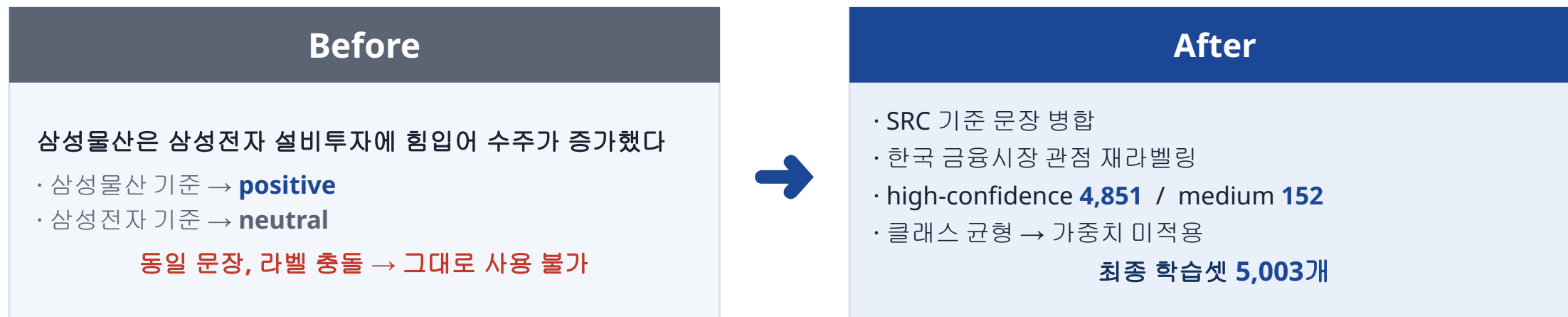
Model	Params	Zero-shot	Standalone	KD
DistilBERT	66M	0.2752	0.9222	0.9071
MiniLM	33M	0.2566	0.8492	0.8913

핵심 인사이트
DistilBERT는 standalone에서 이미 Teacher에 근접 → KD 효과 제한적
MiniLM(33M)은 KD 적용 후 **0.8492 → 0.8913 (+0.042)** 향상
→ 작고 약한 **Student**일수록 **KD**로 보완될 여지 존재

STEP 3

한국어 환경에서 KD 구조 적용

Aspect-level(속성 단위) 구조의 KorFin-ASC 데이터셋을 Sentence-level(단일 문장 기반)로 재구축



클래스 분포



STEP 3

한국어 환경에서 KD 구조 적용

Standalone → KD ($\alpha=0.8$, $T=1.5$, epoch=8)

Model	Params	Standalone	KD
DistilBERT	66M	0.497	0.557
KoELECTRA-small	14M	0.503	0.737
MiniLM	33M	0.332	— (제외)

KoELECTRA-small은 크기가 가장 작지만 **KD 효과가 가장 크게** 나타나 최종 채택

최종 Student

KoELECTRA-small

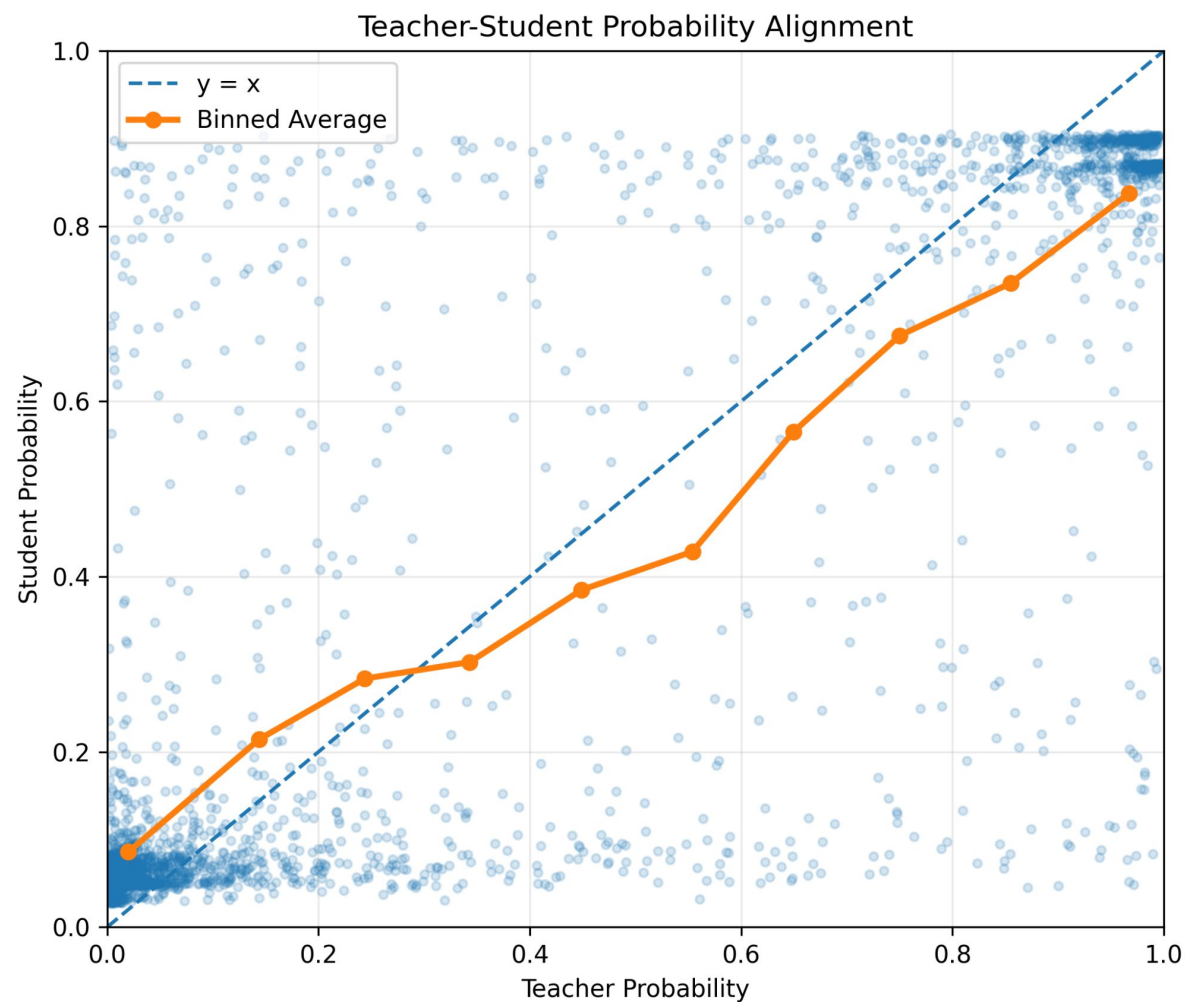
0.737

Macro F1 (KD)

Teacher(0.803)의 **92% 성능 · 1/8 크기**

STEP 4

KD 성공 검증 — 일치도 분석



Teacher Agreement

0.8603

최종 예측 라벨 일치율

KL Divergence

0.2445

확률 분포 차이

Cosine Similarity

0.9085

분포 방향 유사도

STEP 4

영어 / 한국어 비교 분석

영어 — KD 효과 제한적

- 사전학습 데이터가 풍부
- standalone만으로 Teacher에 근접
- Teacher에게서 추가로 배울 여지 작음

DistilBERT 0.9222 → 0.9071

한국어 — KD 효과 뚜렷

- 소형 모델의 사전학습 맥락 부족
- Teacher에게서 배울 값이 큼
- **KD 효과가 크게 발현**

KoELECTRA 0.503 → 0.737 (+0.234)

결론: Standalone 성능이 낮거나 모델 용량이 제한적인 Student일수록 KD를 통해 보완될 여지가 더 크다.

추가 경량화 — Pruning

Method	Sparsity	Macro F1	Retention
KD baseline	0%	0.7892	100%
Pruning only	30%	0.7862	99.62%
Pruning + FT	40%	0.7687	97.41%
Pruning + FT	50%	0.7496	94.99%

30% pruning까지 성능 손실 거의 없음 / 50%도 fine-tuning 적용 시 95% 수준 회복

30% Pruning

99.6%

성능 손실 거의 없음

처리량 (max_len 64)

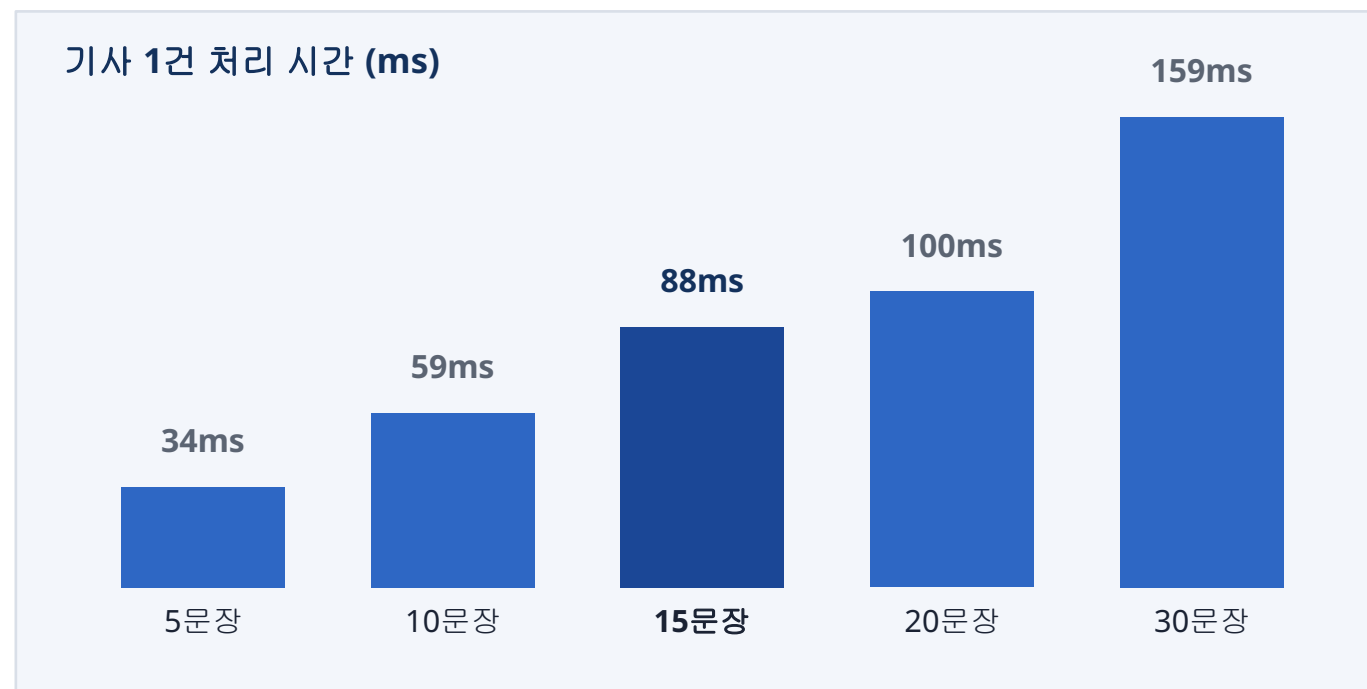
x 1.62

throughput 증가

STEP 6

Chrome 확장 프로그램 구현

네이버 뉴스 클릭 → 문장 분리 → KoELECTRA-small KD 추론 → 기사 감성 집계 / 단일코어 2.8GHz Xeon(최악 조건)



평균 경제 기사 (≈15문장)

0.08~0.09초

클릭 후 사실상 즉시 응답

모델 크기

14M (56MB)

Teacher 110M+ 대비 경량화 이점

헤드라인 뉴스 ⓘ



삼전닉스 2배, '검은 월요일' 직격탄 맞았다...무더기 손실 구간 진입
뉴시스 10+

12



폭락장도 이긴 '깐부' 엔비디아 기대감...SK텔레콤, 2%대 상승
뉴시스 4

4



코스피 폭락에도 환율은 하락...1550원 초중반 등락
이데일리 71

71



美 반도체 쇼크에 30만전자·200만닉스 붕괴[핫스탁]
뉴시스 8

8



현대차, 엔비디아 '피지컬 AI' 동맹에도 9% 하락[핫종목]
뉴스1 35

35



[속보] 젠슨 황 "주가 급락해도 기뻐해야...지금은 할인된 가격에 살 기회"
강원도민일보 50+

31



코스피 '검은 월요일'에 역대급 투매...증권가는 "저가매수 기회"
뉴시스 10+

11



1560원 넘어선 환율...한투證 "과도한 오버슈팅, 금융위기와 달라"
조선비즈 10+

51



피하지 못한 블랙먼데이... 코스피 7400선으로 폭락 후 소폭 반등
조선일보 300+

66



코스피 '블랙먼데이' 공포... "글로벌 펀드들, 하락 방어 강화"
국민일보 50+

20

CONCLUSION

92%

KR-FinBERT 성능 달성

KoELECTRA-small 14M (1/8 크기)

0.08~0.09초

15문장 기사 처리

단일 CPU 코어 실시간 응답

99.6%

30% Pruning 후 성능 유지

추가 경량화 + 처리량 x 1.62

하루 수백 건의 금융 뉴스를 사람이 직접 읽지 않아도, 경량 모델이 실시간 감성 분류로 투자 판단 보조
향후 → 한국어 금융 특화 데이터 확충 · 더 정교한 KD 기법으로 추가 성능 향상 기대

Q & A

경청해 주셔서 감사합니다!