JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 8, pp. 2309-2320, Aug. 2026

Check for updates

LLM 기반 자동 라벨링을 활용한 한국어 금융 뉴스 감정분석 데이터셋 구축과 모델 비교

정 유 리[1] · 김 형 래[2] · 김 성 은[3*]
[1]경북대학교 영어영문학과/경제통상학과 졸업생
[2]경북대학교 컴퓨터공학과 학사과정
[3]경북대학교 기초학문융합연구원 박사후연구원

Dataset Construction and Model Comparison for Korean Financial News Sentiment Analysis: An LLM-Based Automatic Labeling Approach

Yuri Jeong[1] · Hyeongrae Kim[2] · Sungeun Kim[3*]
[1]Alumnus, Department of English Language and Literature and Department of Economics and International Trade, Kyungpook National University, Daegu 41566, Korea
[2]Undergraduate Student, Department of Computer Science and Engineering, Kyungpook National University, Daegu 41566, Korea
[3]Postdoctoral Researcher, Institute for Convergence Research in Basic Sciences, Kyungpook National University, Daegu 41566, Korea

[요 약]

본 연구는 한국어 금융 뉴스 헤드라인 감정분석을 위한 데이터셋을 구축하고, 인코더 기반 모델과 대규모 언어모델의 성능을 비교 및 분석하는 것을 목적으로 한다. 이를 위해 KOSPI 시장 관련 뉴스 헤드라인 데이터를 대상으로 GPT 계열 대규모 언어모델을 활용한 자동 라벨링을 수행하고, Positive, Neutral, Negative의 3-class 감정분석 데이터셋을 구축하였다. 또한 Test set에 대해 경제통상학 전공자 3인의 독립 평가와 평가자 간 일치도 분석을 통해 자동 라벨의 신뢰성을 검증하였다. 구축된 데이터셋을 활용하여 한국어 특화 인코더 모델을 파인튜닝한 결과, 모든 모델이 전반적으로 안정적인 분류 성능을 보였으며, 그 중 KLUE-RoBERTa가 최고 성능을 기록하였다. 한편 Gemini 및 LLaMA 계열 LLM을 Zero-shot 환경에서 평가한 결과, LLM은 추가 학습 없이도 비교적 높은 분류 성능을 보였으나 일부 감정 범주에 편향된 분류 경향을 나타냈다. 본 연구는 한국어 금융 뉴스 감정분석을 위한 자동 라벨링 데이터 구축 방법을 제시하고, 인코더 모델과 LLM에 따른 분류 성능을 실증적으로 분석하였다는 점에 의의가 있으며, 향후 다양한 금융 자연어 처리 연구를 위한 기초 자료로 활용될 수 있을 것으로 기대된다.

[Abstract]

This study aims to construct a dataset for sentiment analysis of Korean financial news headlines and compare the performance of encoder-based models and large language models (LLMs). Using Korea Composite Stock Price Index KOSPI news data, a three-class dataset consisting of Positive, Neutral, and Negative labels was automatically generated through GPT-based pseudo-labeling. Three evaluators independently assessed the appropriateness of the automatic labels in the test set and analyzed inter-annotator agreement. The dataset was then used to train Korean-specific encoder models, while Gemini and LLaMA series models were evaluated in a zero-shot setting. The results showed that encoder models achieved strong classification performance overall, with KLUE-RoBERTa obtaining the best results. Although LLMs demonstrated competitive performance without additional training, they showed imbalanced performance across certain classes. This study presents an automatic labeling approach for Korean financial sentiment analysis and provides empirical insights into the performance characteristics of encoder models and LLMs.

색인어 : 감정분석, 한국어 금융 뉴스, 대규모 언어모델, 자동 라벨링, 인코더 모델
Keyword : Sentiment Analysis, Korean Financial News, LLMs, Automatic Labeling, Encoder-Based Models

http://dx.doi.org/10.9728/dcs.2026.27.8.2309

Received 08 June 2026; **Revised** 16 July 2026
Accepted 05 August 2026

***Corresponding Author; Sungeun Kim**
Tel: +82-53-950-3973
E-mail: emily_kim@knu.ac.kr

Ⅰ. 서 론

금융 시장에서 정보는 자산 가격의 형성과 변동에 중요한 역할을 한다. 그 중 뉴스 텍스트는 투자자들의 인식과 판단에 영향을 미치는 주요 정보원으로 활용된다[1],[2]. 기업 관련 이슈나 경제 상황을 다루는 금융 뉴스는 시장 참여자들의 투자 결정과 판단에 영향을 미칠 수 있다[2],[3]. 실제로 투자자들은 금융 뉴스를 통해 시장 상황과 기업 정보를 파악하고 뉴스의 의견과 감정을 바탕으로 매수·매도 결정을 내리기도 한다[4]. 특히 뉴스 헤드라인을 분석 대상으로 설정하는 것은 실제 투자 환경에서 정보가 활용되는 방식을 반영한다는 점에서 의미가 있으며, 짧은 텍스트에 담긴 제한적인 정보를 기반으로 시장 반응을 추론해야 한다는 점에서 유의미한 연구 과제라고 볼 수 있다.

기존 연구들은 금융·경제 텍스트가 일반적인 도메인과 다른 언어·문맥적 특성을 가지기 때문에 이를 반영한 도메인 특화 감정분석 접근의 필요성을 강조하였다[5],[6]. 대표적으로 Araci는 금융 도메인에 특화된 사전학습 언어모델인 FinBERT를 제안하며 금융 텍스트 분석에서 도메인 특화 사전학습의 중요성을 밝혔다[5]. 또한 Financial PhraseBank는 영어 금융 텍스트에 대해 긍정(Positive), 중립(Neutral), 부정(Negative) 라벨을 부여한 대표적인 데이터셋으로, 금융 감정분석 연구의 주요 벤치마크로 활용되어 왔다[6]. 이러한 연구들은 금융 텍스트의 감정 정보가 시장 반응과 투자 판단에 유용한 신호가 될 수 있음을 보여준다. Tetlock은 뉴스 미디어에 포함된 감정 정보가 시장 반응과 유의미한 관계가 있음을 실증적으로 분석하여 금융 뉴스 감정분석의 활용 가능성을 입증하였다[1]. 그러나 기존의 금융 감정분석 연구와 공개 데이터셋은 주로 영어 금융 텍스트를 중심으로 이루어져 있다는 한계가 있다[7].

한국어 금융 뉴스는 한국어 고유의 언어적 특성으로 인해 영문 기반의 금융 감정분석 연구를 그대로 적용하기 어렵다[7]-[9]. 한국어는 어순, 어미 변화, 문맥 의존성 등에서 영어와 다른 특성을 가지고 있기 때문에 영어 기반 감정 라벨 체계를 한국어 데이터에 직접 적용하기 어렵다는 문제가 있다[7],[8]. 따라서 한국어 금융 텍스트 분석을 위한 데이터 부족은 모델 개발과 성능 향상의 주요 제약 요인으로 작용된다[7],[9],[10]. 현재 한국 금융 시장의 텍스트 어조 분석은 아직 초기 단계에 있으며, 경제 및 경영 분야에서 사용되는 한국어 단어 목록의 부족으로 인해 비정형 텍스트 분석이 어렵다고 보고된다[7]. 그러므로 한국어 금융 뉴스 헤드라인에 특화된 라벨링 체계를 구축하고 실제 시장 반응을 반영한 데이터셋을 마련할 필요가 있다[11]. 이는 한국어의 언어적 특성과 금융 시장의 특수성을 동시에 고려한 금융 뉴스 감정분석 모델 분석을 위한 필수 전제 조건이다.

이러한 배경에서 본 연구는 영문 금융 감정 라벨 체계를 참고하여 한국어 금융 뉴스 헤드라인 데이터셋을 구축하고, 금융 뉴스의 감정을 분류하는 모델을 분석하는 것을 목적으로 한다. 이를 위해 비라벨 상태의 한국어 금융 뉴스 헤드라인에 대해 언어모델이 생성한 예측값을 라벨로 활용하는 자동 라벨링(Pseudo-labeling) 방식을 적용하였다[12]. 최근 대규모 언어모델(LLM; Large Language Model)의 발전으로 자동 라벨링의 활용 범위가 확대되고 있다[13],[14]. 본 연구는 GPT-5 mini(gpt-5-mini-2025-08-07) 기반의 자동 라벨링을 통해 한국어 금융 뉴스 헤드라인 데이터셋을 구축하고, 인간 평가를 통해 라벨의 신뢰성을 검증하였다. 또한 구축된 데이터셋을 기반으로 한국어 특화 인코더 모델과 대규모 언어모델(LLM)의 성능을 비교하고 분석하였다. 인코더 모델은 구축된 데이터셋을 활용한 지도학습을 통해 학습하였으며, LLM은 추가 학습 없이 Zero-shot 환경에서 감정 분류를 수행하였다.

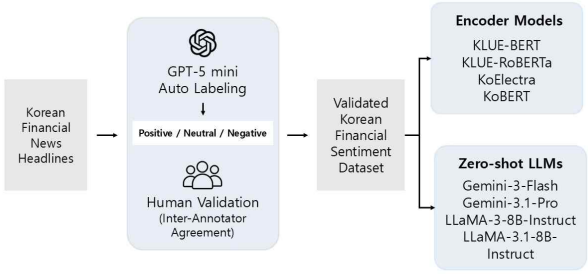

그림 1. 연구 절차
Fig. 1. Research framework

본 연구의 주요 기여는 다음과 같다. 첫째, LLM 기반 자동 라벨링을 활용하여 Positive, Neutral, Negative의 3-class로 구성된 대규모 한국어 금융 뉴스 헤드라인 감정분석 데이터셋을 구축하였다. 둘째, 인간 평가자들이 자동 생성 라벨의 적절성을 독립적으로 검토하고, 평가자 간 일치도를 분석하여 평가자 판단의 일관성을 확인하고 자동 라벨링 데이터셋의 신뢰성을 검토하였다. 셋째, 구축된 데이터셋으로 파인튜닝한 한국어 특화 인코더 모델과 추가 학습 없이 적용한 Zero-shot LLM의 성능 및 분류 특성을 실제 활용 시나리오 관점에서 비교·분석하였다.

Ⅱ. 선행연구

2-1 대규모 언어모델(Large Language Models)

대규모 언어모델(LLM)은 Transformer 구조를 기반으로 대규모 텍스트 데이터를 사전학습하여 다양한 자연어 처리 태스크를 수행하도록 설계된 생성형 언어모델이다[15]. 기존의 사전학습 언어모델이 특정 태스크를 수행하기 위해 추가적인 파인튜닝이 필요했던 것과 달리, LLM은 방대한 파라미

터 규모와 일반화된 언어 이해 능력을 바탕으로 별도의 학습 없이도 문맥 이해와 텍스트 생성 능력에서 뛰어난 성능을 보이며 자연어 처리 분야의 주요 방법론으로 자리잡았다[16].

최근에는 금융 분야에서도 LLM을 활용하는 연구가 활발하게 이루어지고 있다. 대표적으로 FinGPT는 금융 데이터에 특화된 오픈소스 기반의 LLM으로, 금융 감정분석을 비롯한 다양한 금융 자연어 처리 태스크에 적용 가능성을 제시하였다[17]. 이러한 접근은 기존의 지도학습 기반 모델이 가지는 한계를 보완할 수 있을 뿐만 아니라 금융 데이터 추론 및 해석 능력을 강화하는 방향으로 확장되고 있다.

그러나 기존 연구들은 영문 금융 데이터에 집중되어 있어 한국어 금융 텍스트의 적용 가능성은 상대적으로 충분히 검증되지 않은 상황이다. 따라서 본 연구는 한국어 금융 뉴스 헤드라인 데이터에 다양한 언어모델을 적용하여 성능을 비교함으로써 LLM 기반 접근이 한국어 금융 뉴스 환경에서도 효과적으로 활용될 수 있는지 실증적으로 분석하고자 한다.

2-2 LLM 기반 Pseudo-labeling

자연어 처리 분야에서는 라벨 데이터 부족 문제를 해결하기 위해 준지도학습(Semi-supervised Learning) 기반 접근이 활발히 연구되고 있다. 그중 Pseudo-labeling은 비라벨 데이터에 대해 모델의 예측값을 임시 라벨로 부여해 이를 다시 학습에 활용하는 방법으로, 수동 라벨링에 소요되는 비용과 시간을 줄일 수 있다는 장점이 있다[12]. 이후 여러 연구에서 데이터 증강과 신뢰도 기반 학습 전략을 결합하여 Pseudo-labeling의 신뢰성과 학습 성능을 향상시키고자 하였다[18]-[20]. 이러한 선행연구들은 Pseudo-labeling이 라벨링 비용이 큰 도메인에서 유용한 데이터 확장 전략이 될 수 있음을 보여준다.

최근에는 GPT와 같은 대규모 언어모델(LLM)의 등장으로 자동 라벨링 기반의 데이터 구축 방법이 새로운 연구 흐름으로 주목받고 있다. 기존의 Pseudo-labeling이 주로 학습된 분류 모델의 예측 결과를 활용하였다면, 최근 연구에서는 LLM이 보유한 사전학습 지식과 문맥 이해 능력을 활용하여 비라벨 텍스트에 직접 라벨을 생성하는 방식을 제안하고 있다[13],[14]. 별도의 파인튜닝 없이 Zero-shot 및 Few-shot 프롬프팅만으로 비교적 일관된 라벨 품질을 생성할 수 있다는 점에서 LLM 기반 자동 라벨링은 데이터 구축 비용과 시간을 효과적으로 줄일 수 있는 방법으로 평가된다[21]. 또한 일부 연구에서는 LLM이 생성한 라벨이 인간 평가와 비교하여 일정 수준 이상의 일관성과 신뢰성을 확보할 수 있다고 보고하였다[22].

한편 한국어 금융 텍스트 분석 분야에서는 공개적으로 활용할 수 있는 표준화된 데이터셋과 명확한 라벨링 기준이 상대적으로 부족한 상황이다[7]. 금융 뉴스는 일반 도메인과 달리 투자 관점에서의 긍정·부정 정보와 수치 변화, 기업 실적, 시장 전망 등 금융 특화 문맥을 포함하고 있어 라벨링 과정에서 높은 도메인 이해가 요구된다[5]. 이러한 특성으로 인해 대규모 금융 데이터셋을 수작업으로 구축하는 데 상당한 비용과 시간이 소요된다는 한계가 있다.

이에 본 연구는 LLM 기반의 Pseudo-labeling 접근법을 사용하여 한국어 금융 뉴스 헤드라인 데이터셋을 구축하고자 하였다. 기존 금융 감정분석 연구에서 사용된 감정 분류 기준을 참고하여 긍정, 중립, 부정의 3-class 라벨 체계를 설계하였으며, GPT 계열 대규모 언어모델을 활용해 자동 라벨링을 수행하였다. 또한 인간 평가를 통해 생성 라벨의 일관성과 신뢰도를 검토함으로써 LLM 기반 자동 라벨링 방식이 한국어 금융 뉴스 데이터셋 구축에 활용될 수 있는지 살펴보았다.

2-3 금융 감정분석과 특화 언어모델

금융 시장에서 비정형 데이터인 텍스트의 가치를 정량화하려는 시도는 자연어 처리 기술의 발전과 함께 꾸준히 이루어져 왔다[23],[24]. 대표적인 연구인 FinBERT는 일반 뉴스 데이터로 학습된 BERT 기반 모델이 금융 도메인 특유의 언어적 맥락을 충분히 반영하지 못한다는 점에 주목하여 대규모 금융 말뭉치를 추가 사전학습한 모델이다[5]. 해당 연구는 금융 문장을 긍정(Positive), 중립(Neutral), 부정(Negative)의 세 가지 감정으로 분류하는 태스크에서 범용 모델보다 우수한 성능을 입증하였으며, 금융 텍스트 분석에서 도메인 특화 사전학습의 중요성을 보여주었다[5].

이와 같은 금융 감정분석 연구에서 신뢰할 수 있는 라벨링 데이터셋의 구축은 중요한 역할을 한다. 대표적으로 Financial PhraseBank는 금융 관련 배경지식을 가진 평가자들이 뉴스 문장을 긍정, 중립, 부정으로 분류하여 구축한 데이터셋이다[6]. 이 데이터셋은 단순한 감정 표현이 아니라 투자자 관점에서 기업 가치나 주가에 긍정적 또는 부정적 영향을 미칠 수 있는 정보를 기준으로 라벨링되었다는 특징이 있다[6]. 이러한 특성으로 인해 Financial PhraseBank는 금융 감정분석 연구에서 대표적인 벤치마크 데이터셋으로 널리 활용되고 있다[6].

이후 RoBERTa를 비롯한 다양한 Transformer 기반 인코더 모델들이 감정분석 태스크에 적용되면서 성능 개선이 지속적으로 이루어졌다. RoBERTa는 사전학습 전략을 개선하여 기존 BERT 대비 성능을 향상시킨 모델이다[25]. 이러한 인코더 기반 모델은 짧은 텍스트 분류 태스크에서 안정적인 성능을 보이며 감정분석 연구에서 주로 사용된다[26]-[28].

한편 한국어 자연어 처리 성능 향상을 위해 한국어 특성을 반영한 사전학습 언어모델도 활발히 연구되고 있다. KLUE (Korean Language Understanding Evaluation)는 한국어 자연어 이해 능력을 체계적으로 평가하기 위한 벤치마크로 제안되었다[29]. KLUE 환경을 기반으로 학습된 한국어 BERT 및 RoBERTa 계열 모델은 한국어 문맥을 반영한 표현 학습에 활용된다[29]. 또한 KoELECTRA는 generator-discriminator 구조를 활용하여 효율적인 사전학습을 수행하는 모델로, 비교적 적은 연산량으로도 우수한 성능을 보이는

특징이 있다[30]. KoBERT 역시 한국어 말뭉치를 기반으로 사전학습된 모델로, 한국어 문장 구조를 반영한 기본 비교 모델로 사용된다[31]. 이러한 한국어 특화 모델들은 한국어 형태소 구조와 문맥적 특성을 효과적으로 반영함으로써 다양한 한국어 자연어 처리 태스크에 적용되고 있다[32],[33].

본 연구는 이러한 한국어 특화 인코더 모델들을 비교 대상으로 선정하여 금융 뉴스 헤드라인 감정분석에 적합한 모델을 실증적으로 분석하고자 한다. 특히 금융 뉴스 헤드라인과 같이 짧고 문맥 의존성이 높은 텍스트는 일반 도메인과 다른 특성을 가지므로 한국어 특화 언어모델의 적용 가능성을 검증할 필요가 있다. 따라서 본 연구는 한국어 금융 뉴스 환경에서 인코더 기반 모델들이 어떠한 분류 성능과 특성을 보이는지 실증적으로 검토하고자 한다.

Ⅲ. 연구 방법

3-1 학습 데이터 구축

본 연구에서는 KOSPI 시장의 뉴스 데이터를 분석 대상으로 삼았다. 네이버 증권 뉴스(Naver Stock Market News)를 기반으로 수집한 Kaggle의 “Korea Stock Market news title data(KOSPI)”을 활용하였다[34]. 총 700여 개 기업 중 뉴스 발행 빈도와 시장 대표성을 고려하여 상위 50개 기업의 헤드라인을 선정하였다. 최초 수집된 원자료는 총 126,995개의 뉴스 헤드라인으로 구성되며, 헤드라인의 수집 기간은 2022년 4월 16일부터 2023년 8월 12일까지이다[34]. 이 중 헤드라인의 결측값과 빈 문자열, 비정상 라벨과 중복된 헤드라인-라벨 조합을 제거하여 총 117,134개의 뉴스 헤드라인을 최종 분석에 활용하였다.

수집된 뉴스 헤드라인을 감정분석 데이터셋으로 활용하기 위해, 본 연구에서는 GPT-5 mini(gpt-5-mini-2025-08-07) 모델을 활용하여 자동 라벨링을 수행하였다. 금융 뉴스 헤드라인을 입력으로 받아 각 데이터에 대해 Positive, Neutral, Negative 중 하나의 라벨을 부여하였다. 라벨링 기준은 경제 텍스트의 의미적 방향성을 Positive, Neutral, Negative로 분류한 Malo et al.의 연구를 참고하여 설정하였다[6]. 이 데이터셋은 개인 투자자 관점에서 금융 뉴스 헤드라인에 대해 긍정적, 중립적, 부정적 감정을 정의하고 있으며, 구체적인 라벨의 의미는 다음과 같다.

- Positive(긍정): 해당 헤드라인이 주가에 긍정적인 신호를 줄 가능성이 높은 경우
- Neutral(중립): 해당 헤드라인이 주가의 방향성과 직접적인 관련이 없거나 단순 정보 전달 수준인 경우
- Negative(부정): 해당 헤드라인이 주가에 부정적인 신호를 줄 가능성이 높은 경우

자동 라벨링 결과, 최종 데이터셋은 Positive 80,843개(69%), Neutral 21,975개(19%), Negative 14,316개(12%)로 구성되었다. 전체 데이터에서 Positive 클래스의 비율이 가장 높았으며, Neutral과 Negative 클래스가 그 뒤를 이었다. 전체 데이터는 무작위로 Train, Validation, Test set으로 분할하였으며, 각 데이터셋에서도 Positive, Neutral, Negative의 비율이 원자료의 분포인 약 69%, 19%, 12%와 유사하게 유지되도록 하였다. 또한 동일한 데이터 분할 결과를 재현할 수 있도록 무작위 시드값을 고정하였다.

3-2 파인튜닝

본 연구에서는 구축된 한국어 금융 뉴스 헤드라인 데이터셋의 학습 가능성을 확인하고, 뉴스 감정 분류에 적합한 사전학습 언어모델을 도출하기 위해 인코더 기반 모델의 파인튜닝을 수행하였다. GPT-5 mini(gpt-5-mini-2025-08-07)를 통해 자동 라벨링된 총 117,134개의 데이터 중 93,707개(80.0%)를 모델 학습을 위한 학습 데이터셋(Train set)으로 구성하였다. 나머지 데이터는 학습 과정에서 모델의 성능을 검증하고 과적합 여부를 확인하기 위한 검증 데이터셋(Validation set) 11,713개(10.0%)와 최종 모델의 일반화 성능을 측정하기 위한 평가 데이터셋(Test set) 11,714개(10.0%)로 분할하였다. 모델의 성능은 평가 데이터셋(Test set)의 자동 라벨을 기준으로 성능을 평가하였다.

파인튜닝 실험의 입력 데이터는 한국어 금융 뉴스 헤드라인 텍스트로 구성하였으며, 출력 라벨은 Positive, Negative, Neutral의 3-class 감정 범주로 설정하였다. 모델은 뉴스 헤드라인의 문맥 정보를 기반으로 해당 문장이 긍정, 중립, 부정 중 어떤 감정 범주에 해당하는지 분류하도록 학습되었다.

각 인코더 모델별 공정한 성능 비교를 위해 기본 학습 설정은 동일하게 통제하였다. 그림 2와 같이 최대 5 epoch까지의 Validation Loss 추이를 확인한 결과, 손실은 초기 1 epoch에서 크게 감소하고 1-2 epoch 구간에서 최저점을 기록한 후 증가하였다. 이에 데이터 규모와 과적합 가능성, 모델 간 동일한 비교 조건을 고려하여 epoch는 1로 설정하였다. Learning rate는 1e-5를 사용하였다. Batch size는 16으로 설정하였고, 입력 문장의 최대 길이(Max length)는 뉴스 헤드라인의 평균 길이를 고려하여 64로 제한하였다. 최적화 함수(Optimizer)는 Transformer 계열 모델 학습에 주로 사용되는 AdamW를 적용하였다. 비교 모델로는 한국어 자연어 처리 분야에서 대표적으로 활용되는 사전학습 언어모델인 KLUE-RoBERTa, KLUE-BERT, KoELECTRA, KoBERT를 선정하였다. 이 모델들은 서로 다른 사전학습 방식과 구조를 가지고 있어 한국어 금융 뉴스 감정분석에서 성능 차이를 비교하기에 적합하다.

본 연구의 재현 가능성을 높이기 위해 데이터 전처리, LLM 기반 자동 라벨링, 데이터 분할, 인코더 모델 학습 및 평가에 사용된 코드를 GitHub 저장소에 공개하였다. 해당 저장소 주

소는 다음과 같다: https://github.com/Emily-Kim-NLP/stock-data

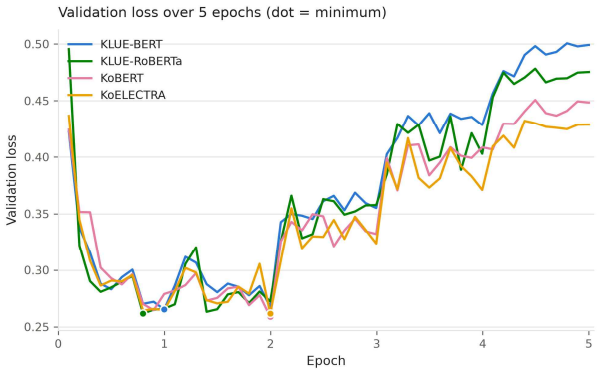

그림 2. Epoch별 validation loss 변화
Fig. 2. Validation loss by Epoch

Ⅳ. 실험 결과 및 분석

4-1 실험 설정

한국어 금융 뉴스 헤드라인 감정분석을 위해 인코더 기반 지도학습 모델과 LLM 기반 Baseline 모델을 함께 구성하였다. 인코더 기반 모델은 구축된 데이터셋을 활용하여 파인튜닝하였으며, LLM은 별도의 추가 학습 없이 Zero-shot 방식으로 감정 분류를 수행하였다. 따라서 본 비교는 두 모델군을 동일한 학습 조건에서 직접 평가하거나 모델 구조의 우열을 판단하기 위한 것이 아니라, 라벨링 데이터를 활용하여 도메인 특화 모델을 학습하는 시나리오와 추가 학습 없이 LLM을 즉시 적용하는 시나리오에서 나타나는 성능과 분류 특성을 비교·분석하는 것을 목적으로 한다. 다만 평가의 일관성을 확보하기 위해 두 모델군에 동일한 Test set과 평가지표를 적용하였다.

LLM 기반 Baseline 모델로는 네 가지 모델 (Gemini-3-Flash, Gemini-3.1-Pro, Llama-3-8B-Instruct와 Llama-3.1-8B-Instruct)을 사용하였다. Gemini 계열은 Google의 폐쇄형 상용 언어모델이며, LLaMA 계열은 Meta에서 공개한 오픈소스 기반의 대규모 언어모델이다.

Gemini-3-Flash는 비교적 경량화된 구조를 기반으로 빠른 응답 속도와 효율적인 추론 성능에 초점을 둔 모델이며, Gemini-3.1-Pro는 보다 심층적인 문맥 이해와 고차원적 논리 추론 성능을 제공하는 상위 모델이다[35]-[37]. Llama-3-8B-Instruct는 다양한 자연어 처리 태스크에서 안정적인 문장 이해 성능을 보이는 모델이며, Llama-3.1-8B-Instruct은 기존 Llama-3-8B-Instruct의 문맥 처리 및 추론 성능을 개선한 모델이다[38],[39]. 본 연구는 상용 LLM과 오픈소스 LLM을 함께 비교함으로써 한국어 금융 뉴스 감정분석에서 모델 계열에 따른 분류 특성을 분석하고자 하였다.

LLM 실험에서는 모든 모델에 동일한 프롬프트를 적용하여 입력된 뉴스 헤드라인에 대해 Positive, Neutral, Negative 중 하나의 감정 라벨을 출력하도록 설계하였다. 프롬프트는 금융 뉴스 감정분석 분류자 역할을 부여하는 System Prompt와 개별 뉴스 헤드라인을 입력하는 User Prompt로 구성하였다. 또한 모델 출력의 무작위성을 최소화하고 일관된 비교 환경을 유지하기 위해 Temperature는 0으로 고정하였다. 사용한 프롬프트는 표 1과 같다.

표 1. LLM 기반 감정 분류에 사용한 프롬프트
Table 1. Prompt used for LLM-based sentiment classification

Prompt type	Prompt content
System prompt	You are a financial news sentiment classifier for Korean stock-market headlines. Classify each headline into exactly one label: pos, neu, or neg. The label pos refers to a headline that indicates a positive financial or business impact. The label neu refers to a headline that is neutral, mixed, factual, or unclear in terms of financial or business impact. The label neg refers to a headline that indicates a negative financial or business impact. Use only the information provided in the headline. Do not infer facts that are not directly supported by the headline. If the headline is ambiguous, mixed, or simply factual, classify it as neu. Return only valid JSON.
User prompt	Classify the sentiment of the following Korean financial news headline. Select only one label among pos, neu, and neg. Headline: {{TITLE}} Return only JSON in the following format: {"label":"pos"}

4-2 평가지표

모델 성능 평가는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score를 활용하였다. Accuracy는 전체 예측 중 모델이 올바르게 분류한 비율을 의미한다. 전체적인 분류 성능을 직관적으로 확인할 수 있다는 장점이 있으나 클래스 불균형이 존재하는 경우 특정 클래스에 편향된 성능을 보일 수 있다. TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative)을 기반으로 계산된다[40].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision은 모델이 Positive로 예측한 것 중 실제 Positive인 비율을 의미한다. 이는 모델의 예측 결과가 얼마나 정확한지 평가하기 위해 사용된다[40].

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall은 실제 Positive인 데이터 중 모델이 Positive로 올바르게 탐지한 비율을 의미한다. 이는 모델이 해당 클래스를 얼마나 놓치지 않고 탐지하는지 평가하는 데 활용된다[40].

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-score는 Precision과 Recall의 조화 평균으로, 두 지표를 균형 있게 반영한다. 클래스 불균형이 존재하는 환경에서 모델의 종합적인 성능을 평가하는 데 유용하다[40].

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

인코더 기반 모델의 학습에는 Cross-Entropy Loss를 사용하였다. Cross-Entropy Loss는 실제 라벨 분포와 모델의 예측 확률 분포 간의 차이를 최소화하도록 학습하는 손실 함수로, 다중 클래스 분류 문제에서 널리 활용된다. C는 클래스 개수, y_i는 실제 라벨(one-hot encoding), $\hat{y_i}$는 모델이 예측한 확률값을 의미한다[41].

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{5}$$

제로샷 프롬프팅(Zero-shot prompting)은 사전에 예제나 데이터를 모델에게 학습시키지 않고 프롬프트 기반 입력만을 활용하여 새로운 작업을 수행하는 방법이다. x는 입력(뉴스 헤드라인), p는 프롬프트, Y는 클래스 집합, $\hat{y}$는 최종 예측 클래스를 의미한다[42].

$$\hat{y} = \arg\max_{y \in Y} P(y|x,p) \tag{6}$$

4-3 인간평가

대규모 데이터셋 구축 과정에서 LLM을 활용한 자동 라벨링은 효율적인 대안이 될 수 있으나, 생성된 라벨이 일관적이고 논리적인 패턴을 갖추고 있지 않는 경우 모델 성능이 저하될 수 있다. 이에 본 연구는 LLM이 생성한 라벨의 품질을 검토하기 위해 인코더 기반 사전 검증과 인간 평가를 병행하였다. 먼저 구축된 데이터셋이 실제 분류 학습에 활용 가능한 패턴을 포함하는지 확인하기 위해, 한국어 자연어 처리에서 널리 활용되는 KLUE-BERT를 기준 인코더 모델로 선정하여 파인튜닝을 수행하였다. 본 절에서 제시한 KLUE-BERT 실험은 최종 모델 간 성능 비교가 아니라, 자동 라벨링 데이터셋의 기본적인 학습 가능성을 확인하기 위한 사전 검증 절차로 수행되었다. 주요 평가지표로는 분류 태스크의 표준인 정확도(Accuracy)와 클래스별 성능을 균형 있게 평가하기 위한 F1-score를 활용하였다. 다음으로 자동 라벨링 결과의 신뢰도를 검증하기 위해 최종 모델 평가에 사용된 Test set의 금융 뉴스 헤드라인 11,714개 전체를 인간 평가 대상으로 선정하였다. 인간평가 결과는 모델 성능 산출이 아니라 자동 라벨의 신뢰도 검증을 위한 보조 절차로 활용하였다. 경제통상학 전공 평가자 3인은 각 뉴스 헤드라인과 LLM이 생성한 Positive, Neutral, Negative 라벨을 함께 확인하고, 해당 자동 생성 라벨이 헤드라인의 금융 감정을 적절하게 반영하는지를 독립적으로 평가하였다. 평가에 앞서 모든 평가자에게 자동 생성 라벨의 적절성을 판단하는 동일한 평가 가이드라인을 전달하였다. 평가자들은 뉴스 헤드라인과 자동 생성 라벨을 비교하여 라벨이 적절하다고 판단한 경우 1(correct), 적절하지 않다고 판단한 경우 0(incorrect)을 부여하였다. 평가자 3인은 서로의 평가 결과를 확인하거나 논의하지 않은 상태에서 모든 항목을 독립적으로 평가하였다. 평가 완료 후 세 평가자의 correct/incorrect 판단을 바탕으로 Gwet's AC1, Average Pairwise Agreement 및 Unanimous Agreement를 산출하였다. 이를 통해 자동 생성 라벨의 적절성에 대한 평가자 간 판단의 일관성과 LLM 기반 자동 라벨링 데이터셋의 신뢰성을 검토하였다.

표 2. KLUE-BERT 검증 결과
Table 2. KLUE-BERT validation results

Model	Dataset	Accuracy	F1
KLUE-BERT	Validation	0.895	0.848
	Test	0.891	0.842

표 3. 인간 평가 라벨 신뢰도 결과
Table 3. Inter-annotator agreement results

Setting	Gwet's AC1	Average pairwise agreement (%)	Unanimous agreement (%)
All three annotators	0.917	92.50	88.76

인코더 기반 모델 검증에서는 표 2와 같이 뉴스 헤드라인을 입력으로 하는 KLUE-BERT 기반의 3-class 분류 모델을 학습하였다. 그 결과, Test set 기준 Accuracy는 0.891, Macro F1-score는 0.842로 높은 수준을 기록하였다. 이는 모델이 입력 데이터로부터 유의미한 분류 기준을 학습하고 있음을 의미하며, LLM 기반 자동 라벨링 데이터가 단순한 무작위 노이즈가 아니라 학습 가능한 패턴을 포함하고 있음을 보여준다.

표 3은 인간 평가를 통한 라벨 신뢰도 검증 결과를 제시한 것이다. 평가자 3인이 동일 데이터에 대해 라벨을 부여한 결과, Average Pairwise Agreement는 92.50%, Unanimous Agreement는 88.76%로 나타나 평가자 간 일치 수준이 높은 것으로 확인되었다. 또한 비대칭적 응답 분포 상황에서도 안정적으로 신뢰도를 측정할 수 있는 Gwet's AC1은 매우 높게 나타났다. 이러한 결과는 평가자들의 판단이 비교적 높은

수준의 일관성을 유지하고 있음을 보여준다.

종합하면, 인코더 기반 모델 검증에서 높은 분류 성능이 확인되었으며, 인간 평가에서도 자동 생성 라벨에 대한 높은 수준의 신뢰성을 보였다. 따라서 본 연구에서 LLM을 활용하여 구축한 자동 라벨링 데이터셋은 안정적인 라벨 품질과 실제 모델 학습에 활용 가능한 수준의 일관성을 갖춘 것으로 판단할 수 있다.

4-4 인코더 모델 평가

한국어 금융 뉴스 헤드라인 감정분석에 적합한 인코더 기반 모델을 알아보기 위해 서로 다른 사전학습 방식과 토크나이징 전략을 사용하는 한국어 사전학습 언어모델들의 성능 비교를 수행하였다. 본 실험에서는 KLUE-RoBERTa, KLUE-BERT, KoELECTRA, KoBERT 총 4개의 인코더 모델을 비교 대상으로 선정하였다. 해당 모델들은 모두 한국어 자연어 처리 분야에서 높은 활용도를 보이는 대표적인 사전학습 언어모델이다.

KLUE-BERT와 KLUE-RoBERTa는 한국어 자연어 이해 벤치마크인 KLUE를 기반으로 학습된 모델로, 다양한 한국어 문맥 이해 성능에서 우수한 결과를 보인다[29]. 특히 KLUE-RoBERTa는 기존 BERT 대비 개선된 사전학습 방식을 적용하여 문장 표현력을 강화한 모델이다[25],[29]. KoELECTRA는 생성자와 판별자를 활용한 사전학습 방식을 적용하여 비교적 적은 연산량으로 효율적인 학습이 가능하다는 특징이 있다[30]. 이러한 특징은 실제 서비스 환경에서의 적용 가능성을 평가하는 데 있어 중요한 기준이 된다. KoBERT는 한국어 코퍼스를 기반으로 사전 학습된 모델로, 한국어 문장 구조를 반영한 언어 표현 학습이 이루어졌다는 점에서 기본 비교 모델로 오랫동안 활용되었다[31].

실험은 모든 모델에 동일한 하이퍼파라미터(Epoch 1, Learning Rate 1e-5, Batch Size 16, Max length 64, Optimizer AdamW)를 적용하여 모델 자체의 성능 차이를 객관적으로 측정할 수 있도록 통제하였다. 평가지표는 Accuracy, Precision, Recall, Macro F1-score를 사용하였다.

실험 결과, 표 4와 같이 KLUE-RoBERTa 모델이 Accuracy 0.897, Macro F1 0.852로 가장 높은 성능을 보이며 전체 모델 중 최적의 성능을 보였다. 또한 KLUE 계열 모델(KLUE-BERT, KLUE-RoBERTa)은 Ko 계열 모델(KoBERT, KoELECTRA)에 비해 비교적 성능이 높게 나타났다. 이러한 차이는 모델 구조와 사전학습 방식의 차이에서 비롯된 결과로 해석할 수 있다.

KLUE 기반 모델은 다양한 한국어 문맥 정보를 반영하도록 학습되어 금융 뉴스와 같은 특수 도메인에서도 상대적으로 우수한 분류 성능을 보인 것으로 판단된다. 특히 KLUE-RoBERTa는 BERT 구조를 기반으로 하면서도 사전학습 전략의 차이를 통해 문장 표현력을 강화한 모델로, 짧은 텍스트인 뉴스 헤드라인에서도 미묘한 의미 차이를 효과적으로 반영할 수 있었던 것으로 보인다.

표 4. 인코더 모델 성능 비교 결과

Table 4. Encoder model comparison results

Model	Dataset	Accuracy	Precision	Recall	Macro F1
KLUE-RoBERTa	Validation	0.902	0.868	0.851	0.859
	Test	0.897	0.862	0.844	0.852
KoELECTRA	Validation	0.893	0.853	0.843	0.847
	Test	0.889	0.847	0.833	0.839
KLUE-BERT	Validation	0.895	0.859	0.840	0.848
	Test	0.891	0.852	0.835	0.842
KoBERT	Validation	0.885	0.841	0.827	0.832
	Test	0.879	0.832	0.817	0.823

KoELECTRA 모델의 경우 상대적으로 가벼운 모델임에도 불구하고 Macro F1이 0.839로 KLUE-BERT(Macro F1 0.842)와 유사한 수준으로 나타났다. 이는 뉴스 헤드라인과 같은 짧은 텍스트에서는 모델의 규모보다 핵심 키워드 간 관계를 효율적으로 파악하는 능력이 중요하게 작용한다는 것을 보여준다. 반면 KoBERT는 상대적으로 가장 낮은 성능(Accuracy 0.879, Macro F1 0.823)을 보였는데, 이는 사전학습 데이터 규모 및 도메인 다양성의 한계에서 기인한 것으로 판단된다.

또한 모델 학습 과정의 안정성과 수렴 양상을 확인하기 위해 체크포인트 스텝(checkpoint step)별 Validation Loss 변화를 분석하였다. 그림 3은 각 인코더 모델의 학습 단계에 따른 Validation Loss 감소 추세를 나타낸 것이다. 전체적으로 모든 모델에서 step이 증가할수록 Validation Loss가 감소하는 경향이 나타났으며, 이는 제한된 학습 설정 내에서 파인튜닝이 비교적 안정적으로 진행되었음을 보여준다. 학습 후반부로 갈수록 Loss 감소 폭이 작아지고 곡선이 완만해지는 형태가 관찰되었는데, 이는 모델들이 주어진 학습 조건 내에서 일정 수준의 수렴(convergence) 경향을 보였음을 의미한다.

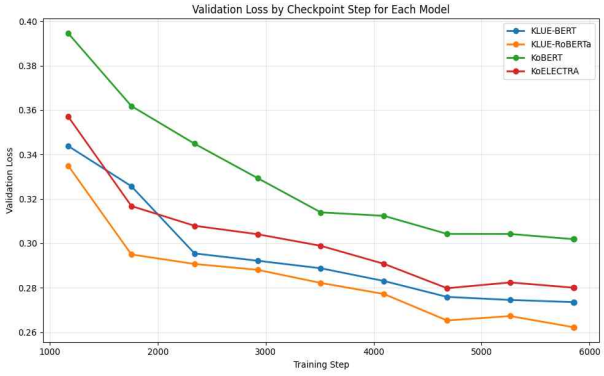

그림 3. 체크포인트별 평가 손실(Validation Loss) 변화

Fig. 3. Changes in validation loss across checkpoints

모델별로 살펴보면, KLUE-RoBERTa가 전체 구간에서 가장 낮은 Validation Loss를 유지하며 가장 안정적인 학습

성능을 보였다. 이는 KLUE-RoBERTa가 금융 뉴스 헤드라인의 문맥적 특징을 가장 효과적으로 학습하고 일반화하였음을 시사한다. KLUE-BERT와 KoELECTRA 역시 비교적 안정적인 감소 추세를 보였다. 반면 KoBERT는 학습이 진행됨에 따라 Loss가 지속적으로 감소하였으나 다른 모델 대비 전반적으로 높은 Loss 값을 유지하였다.

종합적으로 Loss 분석 결과는 모든 모델이 안정적으로 학습되었음을 보여주며, 그중에서도 KLUE-RoBERTa가 가장 우수한 수렴 특성과 일반화 성능을 나타냈다. 성능 평가에서도 KLUE-RoBERTa가 Accuracy와 Macro F1-score 모두 가장 높은 수치를 기록하며 Loss 기반 분석 결과와 부합하는 것을 확인할 수 있었다. 이러한 결과는 금융 뉴스 헤드라인과 같은 짧고 문맥 의존적인 텍스트 분류 태스크에서 사전학습 방식과 문맥 표현 능력이 모델 성능에 중요한 영향을 미칠 수 있음을 보여준다.

4-5 LLM 모델 평가

대규모 언어모델(LLM)이 한국어 금융 뉴스 헤드라인의 감정 분류를 수행할 수 있는지 확인하기 위해 Zero-shot 설정을 적용하여 추가적인 학습 데이터나 Few-shot 혹은 CoT를 제공하지 않은 상태에서 LLM이 도달할 수 있는 성능 수준과 그 한계를 분석하였다.

실험에는 대표적인 거대 언어모델 중 네 가지 모델인 Gemini-3-Flash(gemini-3-flash-preview-12-2025), Gemini-3.1-Pro(gemini-3.1-pro-preview-01-2026), Llama-3-8B-Instruct와 Llama-3.1-8B-Instruct을 사용하였다. 모든 모델은 동일한 프롬프트를 기반으로 뉴스 헤드라인을 입력받아 Positive, Neutral, Negative 중 하나의 라벨을 출력하도록 구성하였으며, 프롬프트 조건을 동일하게 통제하여 모델 간 성능을 비교하였다. 평가지표는 이전 실험과 동일하게 Accuracy, Precision, Recall, Macro F1-score를 사용하였다.

실험 결과, 표 5와 같이 Gemini-3-Flash 모델이Accuracy 0.878, Macro F1 0.842로 Zero-shot 환경에서도 상당히 높은 수준의 분류 성능을 보였다. 이는 추가적인 학습 없이도 대규모 사전학습 과정에서 획득한 언어 이해 능력만으로 일정 수준 이상의 금융 뉴스 감정 분류 태스크를 수행할 수 있음을 보여준다. 반면 동일한 Gemini-3.1-Pro 모델은 상대적으로 낮은 성능(Accuracy 0.775, Macro F1 0.761)을 보였는데, 이는 모델의 규모나 구조적 차이가 항상 성능 향상으로 이어지지 않음을 확인할 수 있다.

LLaMA 시리즈의 경우, Llama-3.1-8B-Instruct은 Accuracy 0.827로 비교적 높은 정확도를 기록하였으나, Macro F1이 0.703으로 전체 비교 모델 중 가장 낮은 수준을 보였다. 이는 일부 감정 클래스에서 성능이 낮게 나타났을 가능성이 있으며, 전체적으로 모든 클래스를 균형 있게 분류하는 데에 어려움이 있다고 해석할 수 있다.

표 5. 모델별 성능 비교
Table 5. Model performance comparison

Model	Accuracy	Precision	Recall	Macro F1
Gemini-3-Flash	0.878	0.835	0.850	0.842
Gemini-3.1-Pro	0.775	0.770	0.800	0.761
Llama-3-8B-Instruct	0.822	0.741	0.742	0.737
Llama-3.1-8B-Instruct	0.827	0.765	0.715	0.703

Zero-shot 환경에서는 모델이 별도의 추가 학습 없이 추론을 수행하기 때문에 특정 클래스에 대한 오분류가 두드러지게 나타났다. 따라서 종합적으로 볼 때 Gemini-3-Flash 모델이 다른 모델에 비해 상대적으로 균형 잡힌 오류 분포를 보였으며, 가장 안정적인 분류 성능을 가지고 있는 것으로 판단된다.

4-6 모델별 사례 분석

모델별 분류 특성을 보다 구체적으로 분석하기 위해 사례 분석을 수행하였다. 인코더 기반 모델과 LLM 모델 간 성능 차이가 나타난 대표 사례를 중심으로 금융 뉴스 헤드라인에 포함된 문맥적 표현과 감정 판단 방식의 차이를 분석하였다.

1) 인코더 기반 모델이 LLM보다 우수한 성능을 보인 사례

본 실험에서 인코더 기반 모델 중 가장 우수한 성능을 기록한 KLUE-RoBERTa와 KLUE-BERT는 Zero-shot 조건에서 평가된 대부분의 LLM 모델을 능가하는 분류 성능을 보였다.

이러한 결과는 금융 뉴스 헤드라인의 제한된 길이 내에서 주요 정보를 압축하여 전달하는 특성과 인코더 기반 지도학습(Fine-tuning)의 효과가 결합된 것으로 해석할 수 있다.

표 6에 있는 예시처럼 '인상', '증가', '하향', '침체'와 같은 표현의 단어들은 문맥에 따라 기업의 실적이나 주가 방향성을 나타내는 핵심 신호로 작용한다. KLUE-RoBERTa와 KLUE-BERT는 한국어 전용 코퍼스를 바탕으로 사전 학습

표 6. 인코더 기반 모델의 우수 분류 사례
Table 6. Correct classification cases by encoder models

Headline	Label	KLUE-RoBERTa	Gemini-3-Flash
KEPCO Expected to Narrow Losses on Possible Additional Utility Rate Hikes	Positive	Positive	Negative
KB Securities Cuts E-Mart Target Price by 21% on Earnings Concerns	Negative	Negative	Positive
Shinsegae International's JAJU Sees 168% Sales Growth for Cooling Fabric Line "JAJU Air"	Positive	Positive	Negative
Daol Investment Cuts Dio's Rating to "Neutral" Amid Real Estate Market Slowdown	Negative	Negative	Positive

된 상태에서 구축된 금융 데이터셋의 분류 패턴을 추가적으로 지도학습을 수행하였기에 금융 도메인 특유의 압축된 표현과 미묘한 어조 변화를 효과적으로 포착할 수 있었던 것으로 보인다.

그 결과, KLUE-RoBERTa(Accuracy 0.897)는 추가 학습 없이 추론을 수행한 Zero-shot 환경의 Gemini-3-Flash(Accuracy 0.878)보다 높은 성능을 기록하였다. KLUE-RoBERTa는 전체 LLM 중 가장 높은 성능을 기록한 Gemini-3-Flash보다도 Accuracy 기준 약 1.9%p 높은 성능을 보였으며, 두 번째로 높은 성능을 기록한 Llama-3.1-8B-Instruct보다 약 7.0%p 이상의 성능 차이를 나타냈다. 이는 추가 학습 없이 Zero-shot으로 적용된 LLM 대비, 특정 도메인 데이터에 파인튜닝을 거친 인코더 기반 모델이 전용 분류 태스크에서 보다 정교하고 안정적인 추론을 수행할 수 있음을 시사한다.

2) 인코더 기반 모델이 LLM보다 낮은 성능을 보인 사례

인코더 기반 모델 중 상대적으로 낮은 성능을 기록한 KoBERT와 KoELECTRA의 경우, LLM 계열의 최상위 모델인 Gemini-3-Flash에 비해 분류 성능이 낮거나 유사한 수준에 머무르는 한계를 보였다. 이는 인코더 기반 모델이 학습 데이터 내 패턴을 중심으로 분류를 수행하는 특성이 강한 반면, LLM은 대규모 사전학습 과정에서 축적한 일반 상식과 문맥 추론 능력을 활용할 수 있다는 차이에서 비롯된 결과로 해석할 수 있다.

표 7에 있는 예시처럼 '적자 축소 기대', '영업손실 축소', '적자 터널 벗어날 것'과 같은 부정적 표현과 긍정적인 전망이 동시에 포함된 문장에서 모델 간 차이가 나타났다. 인코더 모델은 부정적 키워드에 상대적으로 크게 영향을 받아 특정 키워드 중심의 분류 경향을 보였다. 반면 LLM은 문장 전체의 핵심 의미가 회복 신호에 있다는 점을 반영하여 보다 문맥 중심적인 추론을 수행하였다.

그 결과, KoBERT는 Gemini-3-Flash에서 분류 성능이 좋지 않은 부분도 있었다. 이전 결과에서도 Accuracy 0.879, Macro F1 0.823으로 모든 인코더 모델 중 가장 저조한 성적을 기록하였다. 이와 대조적으로 Gemini-3-Flash는 추가적인 파인튜닝이 없는 Zero-shot 환경에서도 Accuracy 0.878, Macro F1 0.842라는 우수한 성능을 기록하였다.

표 7. 인코더 기반 모델의 오분류 사례
Table 7. Misclassification cases by encoder models

Headline	Label	KoBERT	Gemini-3-Flash
Kakao Healthcare Expected to Narrow Losses	Positive	Negative	Positive
LG H&H Expected to See Demand Recovery in Q2	Positive	Negative	Positive
SK hynix Expected to Reduce Operating Loss in Q2, Entering Recovery Phase	Positive	Negative	Positive
Lotte Chemical Expected to Exit Loss-Making Streak After Six Quarters	Positive	Negative	Positive

종합하면, LLM은 추가적인 학습 없이도 다양한 산업 및 경제 이슈에 대한 광범위한 사전지식을 활용할 수 있다는 점에서 강점을 보였다. 금융 뉴스는 단순 감정 표현보다 정책 변화, 시장 전망과 같은 간접적 의미를 포함하는 경우가 많기 때문에 이러한 배경지식 기반의 추론 능력은 일부 사례에서 인코더 모델보다 더 효과적으로 작용할 수 있다.

따라서 인코더 기반 모델은 금융 뉴스 데이터에 대한 파인튜닝을 통해 도메인 특화 표현과 감정 패턴을 안정적으로 학습하여 전반적으로 우수한 성능을 보였지만, 일부 사례에서 경제·금융에 대한 일반 상식과 문맥 추론 능력을 활용하는 LLM이 더 적절한 판단을 수행하기도 한다는 점이 확인되었다. 이러한 결과는 금융 감정분석에서 인코더 모델과 LLM이 서로 다른 강점을 가진다는 점을 보여준다.

V. 결 론

본 연구는 한국어 금융 뉴스 헤드라인을 활용하여 주가 흐름과 관련된 감정을 분류할 수 있는 데이터셋을 구축하고, 다양한 언어모델의 성능을 분석하여 한국어 금융 뉴스 감정분석에 적합한 모델을 살펴보고자 수행되었다. 이를 위해 LLM 기반 자동 라벨링과 인간 평가를 결합하여 데이터셋의 일관성과 학습 가능성을 검토하였다. 또한 인코더 기반 모델을 도메인 데이터로 파인튜닝하는 시나리오와 LLM을 추가 학습 없이 Zero-shot 방식으로 적용하는 시나리오에서 나타나는 성능과 분류 특성을 비교·분석하였다.

인코더 기반 지도학습 모델 비교 실험에서는 KLUE-RoBERTa가 가장 우수한 성능을 보였다. 이는 한국어 금융 뉴스와 같은 짧고 문맥 의존성이 높은 텍스트 환경에서 한국어 문맥 표현 능력과 사전 학습 전략이 모델 성능에 중요한 영향을 미친다는 점을 의미한다. 반면 Zero-shot 기반 LLM 실험에서는 Gemini-3-Flash가 가장 높은 성능을 기록하였다. 그러나 일부 LLM 모델은 특정 감정 클래스에 대한 편향이나 불균형한 분류 성향을 보이며 F1이 상대적으로 낮게 나타났다. 이는 한국어 금융 뉴스 감정분석에서 Zero-shot 환경의 LLM만으로는 모든 감정 클래스를 균형 있게 분류하는 데 한계가 있을 수 있음을 보여준다. 반면 도메인 데이터 기반 파인튜닝을 수행한 인코더 모델은 전반적으로 보다 안정적이고 균형 잡힌 성능을 보였다.

추가적으로 사례 연구를 분석한 결과, 인코더 기반 모델은 금융 데이터셋에 대한 파인튜닝을 통해 도메인 특화 표현을 안정적으로 학습하는 경향을 보였다. LLM은 경제·금융에 대한 일반 상식과 문맥 추론 능력을 활용하여 복합 문맥 구조에서 인코더 기반 모델보다 더 적절한 판단을 수행하는 경우도

확인되었다. 이러한 결과는 금융 감정분석에서 인코더 기반 지도학습 모델과 LLM이 서로 다른 강점을 가진다는 것을 시사한다.

종합하면, 본 연구는 LLM 기반 자동 라벨링으로 구축한 한국어 금융 뉴스 감정분석 데이터셋의 학습 가능성과 라벨 일관성을 인간 평가와 인코더 모델 실험을 통해 검토하였으며, 인코더 기반 모델과 LLM의 분류 특성을 비교하였다. 이에 따라 본 연구는 향후 금융 감정분석, 뉴스 기반 투자 분석, 금융 특화 언어모델 개발 등 다양한 금융 자연어 처리 연구를 위한 기초 자료로 활용될 수 있을 것으로 기대된다.

Ⅵ. 연구의 한계와 향후 연구 방향

본 연구는 LLM 기반 자동 라벨링을 활용하여 한국어 금융 뉴스 감정분석 데이터셋을 구축하고 다양한 언어모델의 성능을 비교하였다는 점에서 의의가 있으나, 몇 가지 한계가 존재한다. 금융 뉴스 헤드라인은 짧고 압축적인 표현으로 구성되어 있어 명확한 감정 신호가 드러나지 않는 경우가 많으며, 이는 클래스 간 경계를 모호하게 만들 수 있다. 실제로 Zero-shot 기반 LLM 실험에서 일부 모델이 특정 감정 클래스에 편향된 예측 경향을 보였다. 이러한 결과는 금융 뉴스 데이터의 표현 특성, 라벨 분포, 모델별 추론 방식에 따라 모델 성능이 달라질 수 있음을 보여준다.

본 연구는 한국어 금융 뉴스 헤드라인의 감정 라벨링 품질과 모델별 분류 성능 비교에 초점을 두었기 때문에, 감정분석 결과와 실제 금융 시장 지표 간의 관계 분석은 후속 연구 과제로 남는다. 향후 연구에서는 분류된 감정 정보가 주가 변동성, 수익률, 거래량 등 시장 지표와 어떠한 상관관계를 가지는지 실증적으로 분석할 필요가 있다.

아울러 LLM 비교 실험은 단일 프롬프트 기반의 Zero-shot 환경에서 수행되었기 때문에 다양한 프롬프트 설계 방식이나 추론 전략에 따른 성능 차이를 종합적으로 분석하기에는 범위가 제한적이었다. 또한 비교 대상 LLM이 Gemini 계열과 LLaMA 계열로 제한되었다는 점에서, 본 연구의 결과를 모든 LLM 계열로 일반화하기에는 한계가 있다. 따라서 Few-shot 설정, 프롬프트 최적화, 다양한 LLM 계열의 추가 비교 등 후속 분석이 필요하다.

향후 연구에서는 금융 뉴스 데이터의 표현 특성과 분포 문제를 고려한 데이터 보완 전략과 인간 평가를 결합한 보다 정교한 라벨링 구축이 필요하다. 금융 감정분석의 궁극적인 활용 목적 중 하나가 주가 방향성 예측인 만큼 본 연구에서 구축한 데이터셋과 파인튜닝 모델을 활용하여 실제 투자 환경에서 뉴스 감정 정보와 시장 수익률 및 변동성 간의 관계를 실증적으로 분석하는 것 역시 중요한 후속 연구 과제가 될 것이다.

감사의 글

본 연구는 교육부 2026년 국립대학육성사업 연구비의 지원에 의하여 이루어진 연구로서, 관계부처에 감사드립니다.

참고문헌

[1] P. C. Tetlock, “Giving Content to Investor Sentiment: The Role of Media in the Stock Market,” *The Journal of Finance*, Vol. 62, No. 3, pp. 1139-1168, June 2007. https://doi.org/10.1111/j.1540-6261.2007.01232.x

[2] D. K. Kirange and R. R. Deshmukh, “Sentiment Analysis of News Headlines for Stock Price Prediction,” *Composoft: An International Journal of Advanced Computer Technology*, Vol. 5, No. 3, pp. 2080-2084, March 2016.

[3] A. Hassanzadeh, A. Razavi, and R. Asadi, “Stock Market Prediction Using Daily News Headlines,” SSRN, 2020. https://doi.org/10.2139/ssrn.3685530

[4] T. L. Im, P. W. San, C. K. On, R. Alfred, and P. Anthony, “Impact of Financial News Headline and Content to Market Sentiment,” *International Journal of Machine Learning and Computing*, Vol. 4, No. 3, pp. 237-242, June 2014. https://doi.org/10.7763/IJMLC.2014.V4.418

[5] D. Araci, “FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models,” arXiv:1908.10063, August 2019. https://doi.org/10.48550/arXiv.1908.10063

[6] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, “Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts,” *Journal of the Association for Information Science and Technology*, Vol. 65, No. 4, pp. 782-796, April 2014. https://doi.org/10.1002/asi.23062

[7] Y. S. Kim and S. W. Cho, “Korean Text Analysis and Applications: Unstructured Text Analysis of Securities Registration Statements Using Machine Learning,” *Korean Journal of Financial Studies*, Vol. 48, No. 2, pp. 215-235, April 2019. https://doi.org/10.26845/KJFS.2019.04.48.2.215

[8] S. Park, Y. Jang, Y. Kang, H. Kang, and H. Kim, “A Study of Corpus Annotation for Aspect Based Sentiment Analysis of Korean Financial Texts,” in *Proceedings of the Annual Conference on Human and Language Technology*, Gyeongju, pp. 232-237, 2022.

[9] Y. Hwang, S. Jung, H. Lee, and S. Yu, “TWICE: What Advantages Can Low-Resource Domain-Specific Embedding Model Bring? – A Case Study on Korea Financial Texts,” arXiv:2502.07131, February 2025. https://doi.org/10.48550/arXiv.2502.07131

[10] B. Hwang, S. Lim, T. Kim, Y. Geun, S. Bang, S. Park, ...

and K. Lim, “KFinEval-Pilot: A Comprehensive Benchmark Suite for Korean Financial Language Understanding,” arXiv:2504.13216, April 2025. https://doi.org/10.48550/arXiv.2504.13216

[11] B. Seo, Y. Lee, and H. Cho, Machine-Learning-Based News Sentiment Index (NSI) of Korea, Bank of Korea, Seoul, BOK Working Paper No. 2022-15, 2022.

[12] D.-H. Lee, “Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks,” in *Proceedings of the ICML 2013 Workshop on Challenges in Representation Learning*, Atlanta: GA, pp. 1-6, June 2013.

[13] F. Gilardi, M. Alizadeh, and M. Kubli, “ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks,” *Proceedings of the National Academy of Sciences*, Vol. 120, No. 30, e2305016120, July 2023. https://doi.org/10.1073/pnas.2305016120

[14] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, ... and X. Shen, “Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, pp. 5085-5109, December 2022. https://doi.org/10.18653/v1/2022.emnlp-main.340

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, “Attention is All You Need,” *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998-6008, December 2017. https://doi.org/10.48550/arXiv.1706.03762

[16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, ... and D. Amodei, “Language Models are Few-Shot Learners,” *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877-1901, December 2020. https://doi.org/10.48550/arXiv.2005.14165

[17] N. Wang, H. Yang, and C. D. Wang, “FinGPT: Instruction Tuning Benchmark for Open-Source Large Language Models in Financial Datasets,” arXiv:2310.04793, October 2023. https://doi.org/10.48550/arXiv.2310.04793

[18] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, “Self-Training with Noisy Student Improves ImageNet Classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle: WA, pp. 10687-10698, June 2020. https://doi.org/10.1109/CVPR42600.2020.01070

[19] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, ... and C.-L. Li, “FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence,” *Advances in Neural Information Processing Systems*, Vol. 33, pp. 596-608, December 2020.

[20] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, “ReMixMatch: Semi-supervised Learning with Distribution Alignment and Augmentation Anchoring,” arXiv:1911.09785, November 2019. https://doi.org/10.48550/arXiv.1911.09785

[21] B. Ding, C. Qin, L. Liu, Y. K. Chia, B. Li, S. Joty, and L. Bing “Is GPT-3 a Good Data Annotator?” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 11173-11195, July 2023. https://doi.org/10.18653/v1/2023.acl-long.626

[22] P. Bansal and A. Sharma, “Large Language Models as Annotators: Enhancing Generalization of NLP Models at Minimal Cost,” arXiv:2306.15766, June 2023. https://doi.org/10.48550/arXiv.2306.15766

[23] O. Shobayo, S. Adeyemi-Longe, O. Popoola, and B. Ogunleye, “Innovative Sentiment Analysis and Prediction of Stock Price Using FinBERT, GPT-4 and Logistic Regression: A Data-Driven Approach,” *Big Data and Cognitive Computing*, Vol. 8, No. 11, 143, November 2024. https://doi.org/10.3390/bdcc8110143

[24] C. H. Joo, Stock Price Prediction and Backtesting Using Financial Text Sentiment Analysis: Application of VADER, FinBERT, and Fine-Tuned FinBERT, Master’s Thesis, Konkuk University, Seoul, 2024.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, ... and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692, July 2019. https://doi.org/10.48550/arXiv.1907.11692

[26] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to Fine-Tune BERT for Text Classification?,” in *Proceedings of the 18th China National Conference on Chinese Computational Linguistics,* Kunming, China, pp. 194-206, October 2019. https://doi.org/10.1007/978-3-030-32381-3_16

[27] T. Bao, N. Ren, R. Luo, B. Wang, G. Shen, and T. Guo, “A BERT-Based Hybrid Short Text Classification Model Incorporating CNN and Attention-Based BiGRU,” *Journal of Organizational and End User Computing*, Vol. 33, No. 6, pp. 1-21, 2021. https://doi.org/10.4018/JOEUC.294580

[28] U. Hasanah, A. M. Soleh, C. Suhaeni, and A. Fitrianto, “Evaluating RoBERTa and GPT-Based Models for SDG Multiclass Text Classification Across Different Document Lengths,” *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, Vol. 20, No. 3, pp. 2645-2664, 2026.

[29] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, ... and K. Cho, “KLUE: Korean Language Understanding Evaluation,” arXiv:2105.09680, May 2021. https://doi.org/

10.48550/arXiv.2105.09680
[30] monologg. KoELECTRA [Internet]. Available: https://github.com/monologg/KoELECTRA.
[31] SKTBrain. KoBERT [Internet]. Available: https://github.com/SKTBrain/KoBERT.
[32] Y. Chen, K. Lim, and J. Park, “Korean Named Entity Recognition Based on Language-Specific Features,” *Natural Language Engineering*, Vol. 30, No. 3, pp. 625-649, 2024. https://doi.org/10.1017/S1351324923000311
[33] E. Lee, C. Lee, and S. Ahn, “Comparative Study of Multiclass Text Classification in Research Proposals Using Pretrained Language Models,” *Applied Sciences*, Vol. 12, No. 9, 4522, May 2022. https://doi.org/10.3390/app12094522
[34] Kaggle. Korea Stock Market News Title Data from Naver Stock [Internet]. Available: https://www.kaggle.com/datasets/park123/korean-news-title-datafrom-naver-stock.
[35] Gemini Team, “Gemini: A Family of Highly Capable Multimodal Models,” arXiv:2312.11805, 2023. https://doi.org/10.48550/arXiv.2312.11805
[36] Gemini Team, “Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context,” arXiv:2403.05530, 2024. https://doi.org/10.48550/arXiv.2403.05530
[37] Google DeepMind. Gemini API Documentation [Internet]. Available: https://ai.google.dev.
[38] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, ... and Z. Ma, “The Llama 3 Herd of Models,” arXiv:2407.21783, July 2024. https://doi.org/10.48550/arXiv.2407.21783
[39] Meta AI, Llama 3.1 Model Card, Meta AI, Menlo Park, CA, July 2024.
[40] M. Sokolova and G. Lapalme, “A Systematic Analysis of Performance Measures for Classification Tasks,” *Information Processing & Management*, Vol. 45, No. 4, pp. 427-437, July 2009. https://doi.org/10.1016/j.ipm.2009.03.002
[41] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Cambridge, MA: MIT Press, 2016.
[42] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, ... and D. Amodei, “Language Models Are Few-Shot Learners,” *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877-1901, 2020. https://doi.org/10.48550/arXiv.2005.14165

정유리(Yuri Jeong)

2026년 : 경북대학교 영어영문학과, 경제통상학과 (학부)

2021년~2026년: 경북대학교 영어영문학과, 경제통상학과 학사
※관심분야 : 금융 데이터 분석, 금융 감정분석, 자연어처리 등

김형래(Hyeongrae Kim)

2026년 : 경북대학교 컴퓨터학부 졸업예정 (학부)

2022년~2026년: 경북대학교 컴퓨터학부 학사
※관심분야 : 자연어처리, LLM, Information Retrieval

김성은(Sungeun Kim)

2021년 : 경북대학교 대학원 영어영문학과 (석사)
2025년 : 경북대학교 대학원 영어영문학과 (박사)

2021년~2025년: 경북대학교 영어영문학과 박사
2025년~2026년: 부산외국어대학교 영어학부 강의초빙교수, 경북대학교 디지털인문공학연구소 객원연구원
2026년~현 재: 경북대학교 기초학문융합연구원 박사후연구원
※관심분야 : 인공지능 활용 교육, 영어교육, 디지털 인문학, 프롬프트 엔지니어링, 자연어처리 등