언어 모델을 활용한 온라인 커뮤니티 정서와 가상화폐 변동성 간의 상관관계 분석

임호준[1], 강승식[2]
[1]국민대학교 소프트웨어학부 학부생
[2]국민대학교 인공지능학부 교수

joken@kookmin.ac.kr, sskang@kookmin.ac.kr

Analysis of the Correlation Between Online Community Sentiment and Cryptocurrency Volatility Using Language Models

Hojun Lim[1], Seungshik Kang[2]
[1]Dept. of Software, Kookmin University
[2]Dept. of Artificial Intelligence, Kookmin University

요 약

뉴스 내용과 증시의 상관관계에 대해서는 다양한 연구가 활발히 진행되었다. 이러한 연구들은 뉴스 제목에 담긴 정보와 증시 변동 사이의 관계를 분석하여 유의미한 결과를 도출하였다. 그에 반해, 직접적으로 드러나는 대중의 반응과 증시의 상관관계에 대해서는 상대적으로 연구가 부족한 실정이다. 본 연구에서는 여러 시간 단위에서 대중들의 반응을 온라인 커뮤니티에서 추출하고, 감정 분석을 통해 수치화 하여 분석한다. 이렇게 수치화 된 감정 데이터가 가상화폐 변동성과 관련이 있는지에 대해 시간 단위 별 상관관계 분석을 통해 알아보고자 한다. 이를 통해 대중의 반응이 가상화폐 시장에 미치는 영향을 실증적으로 분석하고, 가상화폐 시장에 대한 이해도를 높이는 데 기여할 수 있을 것으로 기대한다.

1. 서론

2009 년에 처음 선보인 암호화폐인 비트코인은 수많은 알트코인을 비롯한 모든 가상자산의 대표적인 상징으로 자리 잡고 있다. 이러한 가상화폐의 인기는 탈중앙화의 개념에서 비롯되었다. 2010 년 약 1 달러였던 비트 코인 가격이 2024 년 7 월에는 55,000 달러에 이를 만큼 놀라운 성장을 이뤘으나 가격 변동성 또한 커지면서 비트코인의 가격을 예측하기 위한 연구들이 진행되었다.[1] 또한 최근 몇 년 동안, 텍스트에서 주관적인 정보를 추출하고 분석하기 위해 인공지능 분야를 중심으로 감정 분석 또는 오피니언 마이닝과 관련된 다양한 연구가 진행되었다. 이를 통해 커뮤니티에서 대중들의 반응을 정량화된 수치로 확인할 수 있다.[2] 이 연구에서는 커뮤니티와 같은 포럼 사이트에서 나타나는 일련의 반응들이 투자 결정에 대해 도움이 될 수 있는지 분석하고, 이를 통해 가상화폐 시장에 대한 이해도를 높이는 것에 기여한다.

2. KoBERT 를 활용한 감성 분석 모델

KoBERT 는 한국어 자연어 처리를 위한 BERT (Bidirectional Encoder Representations from Transformers) 모델이다.[3] KoBERT 는 Transformer 아키텍처를 기반으로 한 한국어 사전학습(pretraining) 방법을 사용하여 대량의 텍스트 데이터로 학습되었다. 따라서 한국어 문장에서 단어, 형태소, 구문 등 다양한 언어 단위를 이해하고, 문장의 의미를 파악하는 데 유용하다. 본 연구에서는 문장 단위의 토큰화를 위해 KoBERT 의 SentencePiece 토크나이저를 이용하고, 임베딩된 값을 기준으로 긍정, 부정, 중립을 판단하는 분류 모델을 생성하였다. 모델 학습에 사용할 데이터로는 AI hub 에서 제공하는 감성대화 말뭉치, 한국어 감정 정보가 포함된 단발성 대화 데이터셋 텍스트 데이터를 사용한다. 해당 데이터셋은 SNS 글, 온라인 댓글을 수집한, 감정 분석에 특화된 대략 4 만개의 텍스트 데이터셋이다.

<표 1> 새로운 라벨을 통해 분류를 단순화한 문장의 개수

Emotion	Sentences
긍정	13,376
부정	72,761
중립	10,728

표 1 은 각 감정에 대해 행복은 긍정으로, 슬픔, 분노, 불안, 당황, 공포, 혐오는 부정으로, 중립과 논란은 중립으로 다시 단순화하여 라벨링한 결과이다.

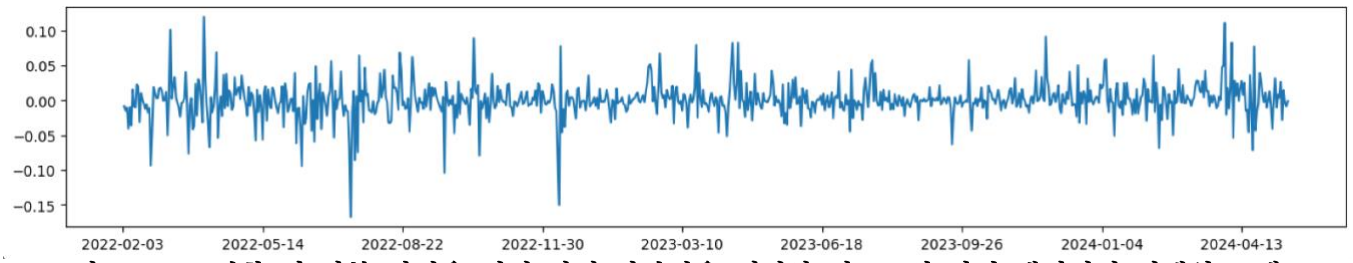

그림 1. 로그 변환 및 차분 과정을 거쳐 자기 상관성을 제거한 비트코인 가격 데이터의 시계열 그래프

그리고 비트코인 변동성에 따른 사람들의 반응의 감정을 수치화하기 위해, 대중의 반응을 텍스트로 확인할 수 있는 데이터를 수집해야 한다. 이런 분석에는 인터넷에 공개 되어있는 여러 커뮤니티나 언론 사이트의 댓글을 참고할 수 있었다. 2 년 6 개월의 기간동안 (2022 년 10 월~2024 년 4 월) 매일 30 개의 게시글을 추출하여 샘플링을 진행하였다. 총 27,000 여개의 게시글 제목을 추출하여 데이터셋을 확보하였다.

<표 2> 게시글 제목 데이터 랜덤 추출 후 평가

제목	Model Output	Label
만우절빔 나오면 웃길듯	[2.3624, -0.3815, -2.2404]	긍정
요즘 어떻게 생각함	[-3.2294, 2.0991, 1.2889]	중립
비트 산사람들 다 도망쳐라	[-3.6801, 1.8766, 2.0699]	부정

표 2 는 앞서 훈련한 모델에 대해 샘플 데이터를 감정별로 추출한 문장에 대해 분류를 수행한 결과이다. 텍스트에 내포된 개별 감정을 잘 분류하는 것을 확인할 수 있다. 해당 모델을 기반으로 22,000 개의 문장에 대해서 분류를 수행하고 상관 관계를 분석하였다.

3. 상관 분석

3.1 데이터 전처리

주가 데이터와 같이, 가상화폐 데이터는 시계열로 두고 봤을 때, 가격 및 거래량 모두 자기상관성을 가지는 특징이 있다.[4] 이를 그대로 예측에 활용하면 예측 범위가 너무 커지게 되어, 분석에 어려움을 겪게 된다. 이를 해결하기 위해 선행적으로 차분을 진행하여 가격 데이터를 변환할 필요가 있다. 또한, 그림 1 과 같이 과거와 현재의 스케일 차이를 해결하기 위해서, 모든 값에 대해 로그 변환을 진행한다. 또한, 더 정확한 분석을 위해 Min-Max Scaling 방식의 정규화를 수행하여 0 ~ 1 로 나타낸다.

3.2 상관계수 분석

위 데이터를 통해, 다양한 시간 단위에 대해서 상관분석을 진행하였다. 여기서는 일, 주, 월, 분기에 대해서, 특정 시간의 감정 분석 결과 및 다음 시간에 대한 감정 분석 데이터 간의 상관관계를 피어슨 상관계수를 활용해 결과를 도출한다. 예를 들어, 지난주의 커뮤니티 반응과 이번주의 비트코인 변동성의 관계를 분석하는 것이다.

<표 3> 시간 단위 별 상관 계수 및 *p-value*

시간 단위(Unit)	상관 계수	p-value
분기 (Q)	0.36	0.35
월 (M)	0.37	0.05
주 (W)	0.19	0.04
일 (D)	0.01	0.76

표 3 은 시간 단위별로 특정 시간 단위에서의 감정 분석 데이터 및 다음 시간 단위에서의 비트코인 가격 데이터 간 상관 계수와 *p-value* 를 나타낸 것이다. 흔히 유의미한 상관 관계를 도출하기 위해서는 *p-value* 는 유의수준인 0.05 보다 작아야 하고, 상관 계수는 0.2 보다 높아야 한다. 분기 단위의 분석에서는 *p-value* 가 0.35 인데 이는 흔히 가설을 기각하는 기준인 0.05 보다 훨씬 높다. 그러나 월별 분석의 *p-value* 는 0.05 로, 유의 수준을 넘지 않는 모습을 보인다. 따라서 분기보다 월별 분석이 상관 관계 분석에서 보다 설득력 있는 자료임을 알 수 있다.

4. 결론

본 연구에서는 SentencePiece 를 활용한 토큰화 기법과 트랜스포머 아키텍처 기반의 KoBERT 사전학습 모델을 사용한 분류기를 활용하여, 커뮤니티 게시글 제목에 따른 감정 분류 실험 및 가상화폐 변동성 데이터 간의 상관 관계를 분석해보았다. 그 결과, 상관계수와 *p-value* 를 같이 확인했을 때 월 단위 분석에서 어느정도 유의미한 양의 상관계수가 있음을 확인할 수 있었다. 상관 계수는 인과 관계가 아닌, 각 변인 사이의 상관 관계만을 확인하는 것이나, 현실 세계에서는 주가 등과 같은 변인을 조작하기가 어려우므로 이 연구를 통해 대중의 반응이 시장에 미치는 영향을 분석함으로써 가상화폐 시장에 대한 이해도를 높이는 데 기여할 수 있을 것으로 기대한다.

Acknowledgement

본 연구는 2024 년도 국민대학교 빅데이터최신기술 수업의 프로젝트 수행 결과이며, 2023 년도 산업통상자원부 ATC+ 사업의 지원을 받았음.

참고문헌

[1] 김선웅, "기계학습 알고리즘을 이용한 알트코인의 가격 예측 성과", 디지털콘텐츠학회논문지, 24(1), 141-151, 10.9728/dcs.2023.24.1.141, 2023.
[2] Pang, B. and L. Lee, "Opinion mining and sentiment analysis". Foundations and Trends in Information Retrieval, 2(1- 2), pp.1- 135, 2008.
[3] Devlin, J., M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", Proceedings of the NAACL-HLT, pp.4171-4186, 2018.
[4] 김진수, "국내증시 일별 거래량의 자기상관성에 대한 고찰", 산업경제연구, 34(4), pp.781-800, 2021.