https://doi.org/10.15813/kmr.2021.22.2.001

텍스트 마이닝과 딥러닝을 활용한 암호화폐 가격 예측 : 한국과 미국시장 비교[1)]

The Prediction of Cryptocurrency on Using Text Mining and Deep Learning Techniques : Comparison of Korean and USA Market

원종관 (Jonggwan Won) 부산대학교[2)]
홍태호 (Taeho Hong) 부산대학교[3)]

〈 국문초록 〉

본 연구에서는 한국과 미국의 대표적인 거래소인 빗썸과 코인베이스의 비트코인 가격을 ARIMA와 순환 신경망(Recurrent Neural Network)을 이용해 예측하고, 이후 각 국가의 뉴스 기사를 이용해 분리 학습에 기반한 separated RNN 모형을 제안한다. separated RNN 모형은 학습 데이터를 가격의 추세 변화 점을 기준으로 분리해 학습시킨 후, 추세 변화 점 별 뉴스 데이터를 활용해 용어 기반 사전을 구축한다. 이후 용어 기반 사전과 평가 데이터 기간의 뉴스 데이터를 이용해 예측할 데이터의 가격 추세 변화 점을 찾아낸 후, 매칭되는 모형을 적용해 예측 결과를 산출한다. 2017년 5월 22일부터 2020년 9월 16일까지의 가격 데이터를 사용해 분석한 결과, 제안된 separated RNN을 이용해 예측한 결과가 한국과 미국의 비트코인 가격 예측 모두에서 순환 신경망(RNN)을 이용해 예측한 결과보다 높은 예측 성과를 보였다. 본 연구는 시계열 예측 기법의 한계를 뉴스 데이터를 이용한 추세 변화 점 탐색을 통해 극복할 수 있고, 성과 향상을 위한 추후 다양한 시계열 예측 기법 및 추세 변화 점 탐색을 위한 다양한 텍스트 마이닝 기법을 적용해볼 필요가 있음을 시사한다.

주제어: 분리 학습, 빈도 기반 텍스트 분석, 순환 신경망, 시계열 분석, 암호화폐

---

1) 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A5A2A01027625).
2) 제1저자, jongkwan1@pusan.ac.kr
3) 교신저자, hongth@pusan.ac.kr

논문접수일 2021년 05월 14일 | 1차 수정 2021년 06월 15일 | 게재확정일 2021년 06월 16일

1. 서론

비트코인은 Nakamo(2008)에 의해 개발된 암호화폐(Crpyto Currency)로, 물리적인 실체가 없으며 블록체인 시스템으로부터 모든 거래 기록이 생성된다. 비트코인의 거래 기록은 임의의 수정이나 삭제가 불가능하며, 따라서 장부 조작을 통한 사기가 거의 불가능하다(Lamothe-Fernández et al., 2020; 권혁준 등, 2018). 또한 비트코인은 은행이나 금융 기관과 같은 중개기관 없이 거래할 수 있고, 개인의 신원과 관련된 정보를 노출시키지 않고 거래할 수 있다는 특징을 가지고 있다(FAUZI et al., 2020; 정윤경 등, 2020.).

이러한 특징을 바탕으로 한 비트코인을 비롯한 암호화폐는 2017년부터 가치가 급상승 하면서 금융경제 부문에서 전 세계적으로 큰 이슈로 주목받았고, 다양한 종류의 블록체인 기술을 활용한 암호화폐들이 등장하고 있다(김선웅, 2021). 뿐만 아니라 학술적으로도 머신러닝 및 딥러닝 기법을 활용한 암호화폐 가격 예측(Chen et al., 2020; Li et al., 2020), 텍스트 마이닝을 활용한 암호화폐 가격 예측(Georgoula, 2015; Yao, 2019), 비트코인 및 암호화폐의 자산적 특성과 미래에 관한 연구 등(Extance, 2015; Luther, 2016; Wolla, 2018) 암호화폐 및 비트코인과 관련된 다양한 연구가 진행되고 있다. 특히 비트코인은 통화수단보다는 주로 주식, 금과 같이 금융 투자 상품으로 분류되어져 왔고(허인, 2019; Cheah & Fry, 2015), 이와 관련해 가격의 변동 요인을 찾는 연구(Georgoula, 2015), 미래 가격을 예측하거나 변동성을 예측하는 연구(Velankar et al., 2018)가 진행되었다. 유사한 특징을 가지고 있는 주식이나 금의 가격을 예측하는 연구들(Tang et al., 2009; Pawar et al., 2019)에서 사용된 통계적 기법인 ARIMA, 머신러닝 기법인 인공 신경망(Neural Networks), 서포트 벡터 머신(SVMs), 딥러닝 기법인 순환 신경망(Recurrent Neural Networks), LSTM, 합성곱 신경망(Convolutional Neural Networks) 등 다양한 기법을 사용해 비트코인의 가격을 예측하는 연구가 진행되어왔다(Azari, 2019). 이와 더불어 뉴스 기사, 트위터 데이터를 텍스트 마이닝을 통해 투자자들의 감성을 추출해 비트코인 가격 예측에 활용한 연구 또한 일부 진행되었다(Corbet et al., 2020; Karalevicius et al.,2019).

비트코인의 가격과 같은 시계열 데이터의 예측에 있어, 장기간의 예측은 오차의 누적, 정보의 부족 등으로 인해 단기간의 예측에 비해 많은 어려움을 가지고 있다(Makridakis, 1994). 관련된 연구에서도 추세가 일정한 단기 가격의 예측 성과는 일반적으로 추세가 변하는 장기 가격의 예측 성과보다 뛰어날 수 있음을 보였는데, Azari(2019)는 ARIMA 기법을 활용해 비트코인 가격을 예측하는 연구에서 ARIMA 기법을 적용해 모형을 구축했다. 연구 결과 추세가 변하지 않는 단기적인 예측에서는 유용성을 보였지만, 추세가 변하는 장기적인 예측에서는 유용성을 보이지 않았다. 비트코인과 같은 금융 투자 상품을 매매함에 있어 추세의 변화를 파악하는 것은 손실을 최소화하고, 수익을 최대화하는 데 있어 중요한데(Abad et al., 2004; Chang, 2012), 이와 관련해 금융 투자 상품 추세의 변화를 예측하는 연구들이 진행되고 있다(Kodama, et al., 2017; Telli & Chen, 2020). 또한 추세 별로 데이터를 분할하고 분할된 데이터 셋 별로 각각 학습, 평가를 하는 연구도 진행되었다. 배성완 & 유정석(2018)은 추세별로 부동산 가격 데이터를 분리한 뒤, 분리된 데이터 셋 별로 학습 모형 구축 및 예측 성과를 평가를 진행한 결과 시기별로 예측 정확도가 다를 수 있음을 보였다. 조보근 등(2020)은 지역 별 아파트 가격지수 예측에 추세 별로 데이터 셋을 분리하여 분리된 데이터 셋 별로 모델을 학습시키고 예측한 결과 특정 지역에서는 단일 데이터 셋을 사용해 예측한 이전 연구들

보다 더 좋은 성과를 보였다. 한편 금융 투자 상품의 추세를 파악하는 방법으로 텍스트 데이터를 활용하는 연구들이 진행되고 있는데, 텍스트 마이닝 기법을 활용해 예측할 데이터와 관련된 배경, 의미, 추세 등을 파악할 수 있다. Hagenau et al.(2013)는 주가를 예측함에 있어 독일과 영국의 뉴스 데이터에 TF-IDF를 적용해 주가의 상승, 하락에 영향을 미치는 긍정, 혹은 부정적인 단어들을 추출한 뒤 변수로 사용했고, 서포트 벡터 머신을 사용한 주가 예측 모형의 성과를 높였다. Mittermayer(2004)은 주가 예측에 있어 일일 뉴스 데이터를 뉴스 자동 처리 시스템, 자동 분류 시스템을 구축해 당일 매매(day-trading)의 수익률을 높일 수 있는 예측 모형을 구축했다.

또한 추세 예측 이외에도 비트코인 가격의 경우 거래소 별, 국가별로 가격에 대한 합의가 이루어지지 않은 상태가 발생하며(Brandvold, 2015), 특히 한국 거래소의 경우 비트코인 가격이 프리미엄의 형태로 다른 국가들의 비트코인 가격보다 높게 나타난 시기들이 다수 존재했고(김효상, 2019), 이러한 점은 비트코인 가격 예측에 있어 추가적인 어려움을 가져다 줄 수 있다.

이전의 내용들을 종합해보면, 추세가 변하는 구간들이 존재하는 장기간의 데이터를 사용하여 가격 예측 모형을 구축할 경우 예측 정확도가 낮아질 수 있다. 하지만 시계열 데이터를 추세 별로 분리한 후, 분리된 구간별 데이터로 학습한 다수의 예측 모형을 구축한다면 단일 데이터로 학습한 단일 예측 모형보다 예측 성과를 향상시킬 수 있다. 또한 새로운 데이터를 예측 할 때 뉴스 데이터를 활용해 예측할 데이터와 유사한 추세를 가지고 있는 학습 데이터를 찾는다면, 추세 별로 다수의 모형들 중 알맞은 모형을 적용할 수 있다. 추가적으로 비트코인의 경우 국가별로 비트코인 가격에 차이가 존재할 수 있다. 본 연구에서는 유사한 특성을 가지고 있거나 유사한 추세를 가지고 있는 부분들로 전체 데이터 셋을 나눈 후, 분리된 데이터 셋 별로 각기 다른 모형을 구축하는 분리 학습 모형을 제시하고, 한국과 미국의 비트코인 가격 예측을 통해 유효성을 증명하고자 한다. 연구를 통해 뉴스 데이터를 기반의 지식을 활용해 비트코인 등 암호화폐 매매나, 유사한 특징을 가진 주가 등의 매매에 유용하게 사용될 수 있다.

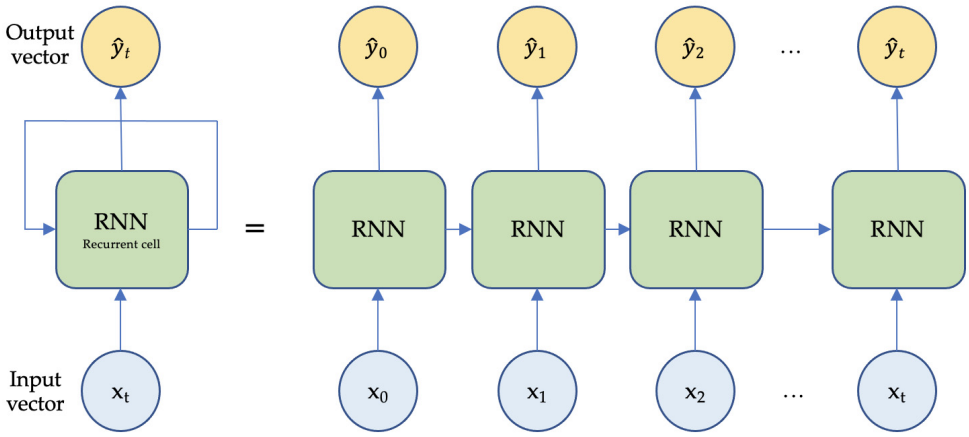
본 논문의 구성은 다음과 같다. 2장에서는 순환 신경망, 빈도 기반 텍스트 분석, 분리 학습(Separate Learning)에 대한 선행연구에 대해 알아본다. 3장에서는 본 연구의 진행 구조인 연구 프레임워크에 대해 살펴본다. 4장에서는 제안된 방법론을 바탕으로 한 실험 과정을 알아보고, 5장에서는 실험 결과를 설명한다. 마지막 6장에서는 결론과 본 연구의 의의 및 한계를 제시한다.

2. 문헌 연구

2.1. 순환 신경망(Recurrent Neural Networks, RNN)

순환 신경망(Recurrent Neural Networks, RNN)은 하나의 입출력 패턴을 가진 인공 신경망 병렬체인 구조를 연결한 형태로, 과거 학습결과를 현재 학습에 사용하는 딥러닝 네트워크이다. 순환 신경망의 구조는 <그림 1>과 같다. 순환 신경망은 그림과 같이 입력층, 은닉층, 출력층의 3단계 구조로 이루어져 있다. 순환 신경망의 특징은 은닉층이 이전 데이터를 참조하도록 서로 연결되어 있다. 즉, 입력값 x_t는 y_t라는 결과값을 출력함과 동시에 다음 출력값인 y_{t+1}에 영향을 미친다.

순환 신경망은 텍스트 생성, 시계열 예측 등의 분야에서 주로 사용되고 있는데, 특히 시계열 예측과 관련



〈그림 1〉 순환 신경망(Recurrent Neural Networks, RNNs) 구조도

해 주가 예측에 빈번히 활용되고 있다(Kraus et al., 2017). Selvin et al.(2017)는 짧은 기간의 주가 예측에 ARIMA, 순환 신경망은, LSTM, 합성곱 신경망 기법을 활용한 모형을 각각 만들어 그 성과를 비교했다. 딥러닝 기법인 순환 신경망, LSTM, 합성곱 신경망의 경우 예측 시기에 따라 서로간의 정확도에 차이가 있었지만, 통계적 기법인 ARIMA 모형보다는 조건과 기간에 상관 없이 높은 예측 정확도를 보일 수 있음을 실증분석을 통해 주장하였다. Pawar et al.(2019)는 LSTM 셀을 사용해 순환 신경망의 장기 의존성 문제를 해결한 사용한 주가 예측 모형을 구축했고, 인공 신경망, 서포트 벡터 머신(Support Vector Machine)을 사용한 주가 예측 모형과 비교해 순환 신경망 기법의 우수성을 밝혔다. 또한 암호화폐의 가격 예측에도 활용되고 있는데, Velankar et al.(2018)는 LSTM 셀을 사용한 순환 신경망을 이용한 예측 모형을 만들었다. 인공 신경망, 서포트 벡터 머신, 의사결정나무 알고리즘을 사용한 모형들보다 예측 성과가 약 97.2%로 가장 높았고, 이를 통해 순환 신경망의 우수성을 증명할 수 있었다. 또한 Felizardo et al.(2019)는 순환 신경망, 서포트 벡터 머신, ARIMA 등 다양한 기법을 적용한 모형들 간 회귀분석 결과 비교를 통해, 장기간의 예측에서는 LSTM셀을 이용한 순환 신경망을 적용한 모형이 다른 기법을 적용한 모형에 비해 더 우수한 성과를 나타낸다는 것을 주장하였다.

2.2. 빈도 기반 텍스트 분석(Frequency based Text Analysis)

문서를 분석하기 위해서는 텍스트를 숫자로 바꿔주는 과정이 필요하다. 이를 가리켜 임베딩(embedding) 혹은 벡터화(vectorization)이라고 한다. 문서를 벡터화하는 방법으로는 Bag of Words(BoW)가 있다. BoW란 문장에 등장하는 단어들의 빈도를 이용하는 방법이다. 단어의 순서와 상관없이, 해당 단어가 문서에 몇 번 나왔는지를 따진다. 다른 방법으로는 TF-IDF(Term Frequency - Inverse Term Frequency)가 있다. TF-IDF는 문서에 포함되어 있는 단어의 중요성에 따라 단어와 문서의 연관성을 계산하는 방법이다. 문서에 등장하는 단어의 상대적인 중요성을 수치로 정규화하여 점수를 계산한다. 아래 수식은 TF-IDF의 수치를 계산하는 수식이다. 단어 빈도인 $tf(t,d)$의 경우, 이 값을 산

출하는 다양한 방법이 존재하지만, 가장 간단한 방법은 수식(1)과 같이 단순히 문서 내에 나타나는 해당 단어의 총 빈도수를 사용하는 것이다. 문서 d 내에서 단어 의 총 빈도를 $f(t,d)$라 할 경우, $tf(t,d)=f(t,d)$로 표현해 산출한다. 역문서 빈도인 $idf(t,D)$는 한 단어가 문서 집합 전체에서 얼마나 공통적으로 나타나는지를 나타내는 값이다. 전체 문서의 수($|D|$)를 해당 단어를 포함한 문서의 수($\log\frac{|D|}{|d\in D:t\in d|}$)로 나눈 뒤 로그를 취하여 얻을 수 있다(Dillon, 1983).

$$tf(t,d)=f(t,d) \quad (1)$$

$$idf(t,D)=\log\frac{|D|}{|d\in D:t\in d|} \quad (2)$$

빈도 기반의 텍스트 분석은 시계열 예측 분야에서 통계적 기법 및 딥러닝 기법과 병행되어, 혹은 단독적으로 사용되고 있는데, Tang et al.(2009)는 주가 예측에 있어 시계열 기법인 SVR(Support Vector Regressor)과 함께 빈도 기반의 텍스트 분석 기법을 적용해 뉴스 기사가 주가에 변동에 대한 추가적인 정보를 제공하는 것을 증명했다. 또한 Mittermayer(2004)는 TF-IDF를 활용해 주가의 상승과 하락에 영향을 미치는 단어들을 선정해 수익률을 높일 수 있는 거래 시스템을 구축했다. 또한 주가 예측 뿐 아니라 비트코인을 포함한 암호화폐 가격 예측에도 사용되고 있는데, Yao et al.(2019)는 TF-IDF를 활용해 뉴스 기사로부터 비트코인 가격에 영향을 주는 단어들을 변수로 추출하고, 이를 서포트 백터 머신, 인공 신경망, 로지스틱 회귀분석 등을 활용한 예측 모형에 변수로 추가하여 예측 성능을 높였다. 또한 Georgoula(2015)는 트위터 데이터의 단어 빈도(term frequency)를 활용해 문장을 Vectorization한 후, 서포트 벡터 머신 기법을 시계열 데이터에 적용해 비트코인의 가격에 영향을 미치는 다양한 변수들 중 하나로 사용해 유용성을 밝혔다.

2.3. 분리 학습(Separated Learning)

본 연구에서 정의하는 분리 학습은 3단계로 이루어져 있다. 첫 번째, 시계열 데이터인 학습 데이터 셋을 유사한 추세를 가지고 있는 부분들로 분리한다. 가격의 추세가 변화하는 부분, 즉 추세 변화 점을 기준으로 데이터 셋을 분리하고 분리된 학습 데이터 셋 별로 따로 시계열 예측 기법을 적용해 각기 다른 모형을 구축한다. 두 번째, 학습 데이터 기간의 뉴스 데이터를 활용해 추세 변화 점 별로 키워드를 정리한다. 세 번째, 평가 데이터 기간의 뉴스 데이터와 정리된 키워드들을 활용해 평가 데이터 기간의 추세 변화 점을 찾고, 찾은 추세 변화 점과 매칭되는 모형을 적용해 가격을 예측한다. 시계열 데이터를 예측할 때는 추세가 일정한 데이터를 학습하고 예측할 경우 정확도가 그렇지 않은 경우에 비해 더 높을 수 있다(Azari, 2019). 또한 추세 별로 데이터를 분리해서 학습한 모형의 경우 그렇지 않은 모형에 비해 예측 정확도를 높일 수 있다(조보근 등, 2020). 따라서 같은 추세를 가지고 있는 데이터들 끼리 분리하는 아이디어에 기반에 학습 기법을 적용하게 된다면 예측 성과가 높아질 수 있다.

분리 학습 아이디어를 적용한 연구는 다양한 분야에서 시도되고 있는데, Lu et al.(2012)은 고객 이탈 예측에서 단순히 전체 데이터 셋을 가지고 예측 모형을 만들지 않고, 고객들을 두 집단으로 나눈 뒤, 특성을 반영한 부스팅 알고리즘을 사용해 예측 성과를 높였다. 또한 시계열 예측 분야에서는 시계열 데이터를 wave filtering 기법을 사용해 추세 부분과 변동부분으로 나눈 뒤, 각각의 데이터에 대해 다른 모형을 구축해 성과를 높인 연구가 존재한다(Joo & Kim, 2015). 배

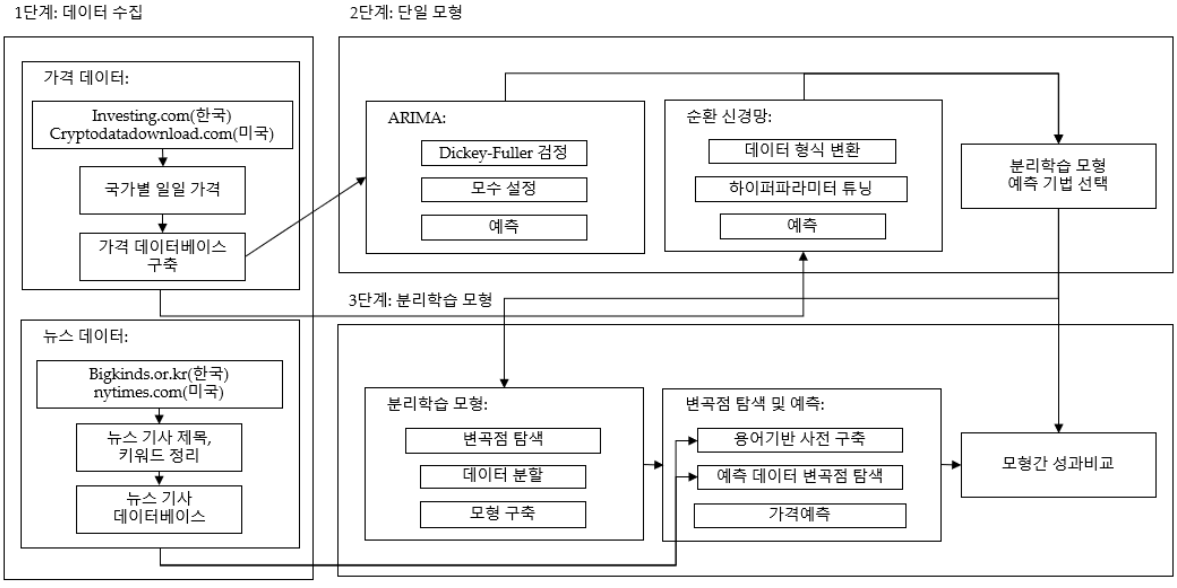
성완 & 유정석(2018)은 시세가 급변하는 시기와 비교적 안정적인 시기로 부동산 가격 데이터를 분리한 뒤, 개별적으로 학습 및 평가를 한 결과 시기별로 예측 정확도가 다를 수 있음을 보였다. 또한 조보근 등(2020)은 지역 별 아파트 가격지수 예측에 지수 추세 변화점과 개입 효과가 일어나는 부분을 분리하여 각 시점별로 모델을 학습 및 예측한 결과 특정 지역에서는 단일 데이터 셋을 이용해 모형을 구축한 이전 연구들 보다 더 높은 성과를 보였다.

3. 연구 프레임워크

<그림 2>는 연구의 프레임워크다. 본 연구에서는 한국과 미국시장의 비트코인 가격 지수를 각각 빗썸(Bithumb) 거래소의 가격 데이터와 코인베이스(Coinbase) 거래소의 가격 데이터를 사용한다. 또한 뉴스 데이터의 영향력을 확인하기 위해 한국 뉴스기사와 미국 뉴스기사 데이터를 사용한다.

연구는 세 단계에 걸쳐서 진행된다. 첫 번째 단계에서는 가격 데이터와 뉴스 데이터를 수집한다. 가격 데이터는 한국 빗썸 거래소의 경우 Investing.com로부터, 미국 코인베이스 거래소의 경우 Cryptodatadownload.com으로 부터 일일 가격 데이터를 다운로드 받는다. 이후 거래기간 통일, 화폐 통일(USD) 등 전처리를 진행한 후 가격 데이터 데이터베이스를 구축한다. 뉴스 데이터는 한국의 경우 빅카인즈(Bigkinds) 사이트에서 제목, 키워드를 다운로드 받아서 일일 별로 단어들을 정리한다. 또한 미국의 경우 Python 3.8과 Newyork times API(Application Programming Interface)를 활용해 기사별 제목, 키워드를 내려 받은 후 마찬가지로 일일별로 단어들을 정리한다.

두 번째 단계에서는 단일 모형들을 ARIMA, 순환신경망 기법을 사용해 구축한다. ARIMA의 경우 Dickey-Fuller 정상성 검사를 통해 차분 횟수를 결정하고, ACF/PACF를 활용해 절단값을 결정해 ARIMA 기법의 모수를 선정하게 된다. 순환 신경망 모형의 경우 python의 keras 라이브러리를 사용해 구축을 하기 때



〈그림 2〉 연구 프레임워크

문에, 모형에 맞게 데이터 형식을 변경해준다. 이어 학습데이터와 검증데이터를 활용해 하이퍼파라미터(Hyperparameter)를 튜닝하고, 튜닝한 하이퍼파라미터를 바탕으로 생성한 모형을 바탕으로 예측값을 생성한다. 추후 단일 모형들은 분리 학습 모형의 우수성을 확인하기 위한 벤치마크로 사용되고, 단일 모형 중 더 우수한 성과를 가지는 기법은 분리 학습 모형에 적용할 기법으로 선정된다.

세 번째 단계에서는 분리 학습 모형을 구축한다. 먼저 학습 데이터의 추세 변화 점들을 선정한다. 이어 점들을 기준으로 데이터 셋을 분리하고, 분리된 데이터 셋 별로 2단계에서 더 우수한 예측 성능을 보인 기법을 적용해 다수의 모형을 구축한다. 이어 추후 예측할 데이터의 추세 변화 점을 찾기 위해 일별 뉴스데이터 키워드 데이터를 활용해 학습 데이터의 추세 변화 점 별 용어 사전을 구축한다. 이어 일별 용어 빈도 분석을 통해 예측할 데이터의 추세 변화 점을 탐색한다. 이후 찾게 된 점과 매칭되는 모형을 적용해 예측 값을 생성한다. 마지막으로 벤치마크 모형인 ARIMA, 순환 신경망과 분리 학습 모형 간 성과 비교를 통해 모형의 우수성과 유효성을 파악한다.

4. 실험

4.1. 1단계: 데이터 수집

데이터 수집은 비트코인 가격 데이터, 뉴스 기사 데이터 두 부분으로 나누어서 진행하게 된다. 실증 분석을 위한 가격 데이터는 2017년 5월 22일부터 2020년 9월 16일까지의 일별 비트코인 종가 데이터이다. 다양한 암호화폐가 존재하지만 비트코인을 분석 대상으로 선택한 이유는 비트코인이 암호화폐 시장에서 시장점유율, 거래량, 시가총액이 가장 높고, 다른 암호화폐에 미치는 영향이 가장 크기 때문이다(Ciaian & Raicaniova, 2018). 한국의 비트코인 가격은 빗썸 거래소의 데이터를, 미국의 비트코인 가격은 코인베이스 거래소 데이터를 수집했다. 다른 거래소의 경우 분석 대상으로 삼기에는 거래량이나 거래 기간이 지나치게 작아 각 국가를 대표하는 거래소라 보기 힘들기 때문이다. 한국 빗썸 거래소의 비트코인은 원화(₩)로 거래되기 때문에 표기 화폐 또한 원화이다. 본 연구에서는 국가별 비트코인 가격을 비교하기 위해 2017년 5월22일에서 2020년 9월 16일까지의 우리은행 일일 환율 데이터를 수집해 달러화($)로 환산했다. 또한 데이터의 학습과 평가를 위해 학습 데이터 80%, 평가 데이터 20%로 나누어 실험을 진행했다.

<표 1>은 각 거래소의 가격 지수에 대한 기초 통계량을 나타낸다. 한국 거래소인 빗썸에서 거래된 비트코인 1개의 평균 가격은 $7718.97로 미국 거래소인 코인베이스의 평균 가격인 $7493.09에 비해 약 3%이상 높았고 최대 가격은 약 20%가까이 높았다. 뿐만 아니라 최소 가격, 표준편차 또한 빗썸에서 더 높은 것으로 나타났다.

한편 예측할 데이터의 추세 변화 점 탐색을 위해 한국 뉴스 기사의 경우 뉴스 빅데이터 분석 시스템인 빅

〈표 1〉 거래소 별 비트코인 가격 기초 통계량(단위: 달러)

거래소	평균	표준편차	최댓값	최솟값
빗썸	7718.9	3306.5	23487.5	1948.0
코인베이스	7493.1	2983.3	19650.0	1910.8
차이(difference)	225.9	794.4	7355.3	37.2

카인즈 에서 다운로드 했고, 미국 뉴스 기사의 경우 뉴욕타임스가 제공하는 API를 통해 일자, 제목, 키워드를 추출했다.

4.2. 2단계: 단일 모형

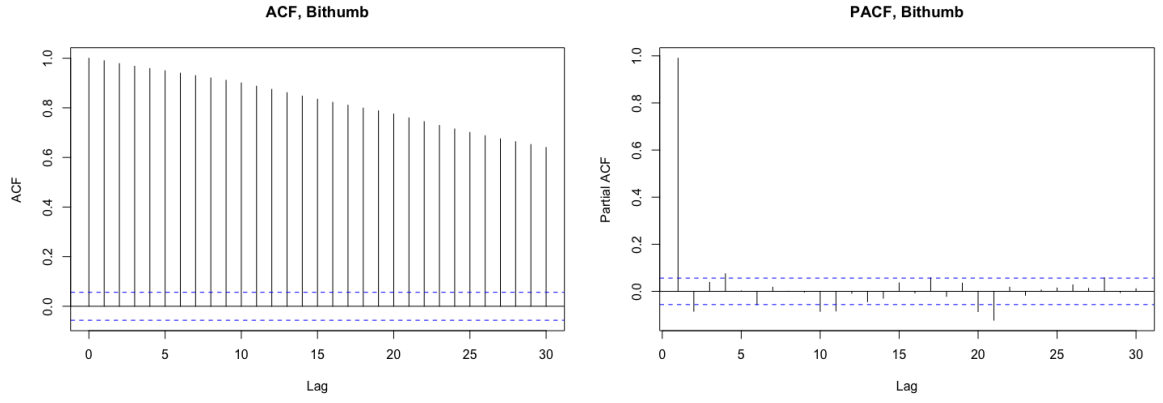
4.2.1. ARIMA 모형

ARIMA 모형을 구축할 때는 모수(parameter) 설정이 성과에 있어서 중요하다. 모수는 차분 횟수인 *d*와 AR 모형의 차수인 *p*, MA 모형의 차수인 *q*로 구성되어 있다. 본 연구에서는 빗썸과 코인베이스의 비트코인 가격 데이터에 대해 각각 Dickey-Fuller의 단위근 검정을 실시해 차분 횟수 *d*를 결정했다. 차분을 해나가며 검정을 실시하고, 검정 결과 *p*값이 0.05보다 작을 경우 정상성을 만족한다고 판단하여 최초 p값이 0.05보다 작은 차분 횟수를 *d*로 결정한다. <표 2>는 Dickey-Fuller

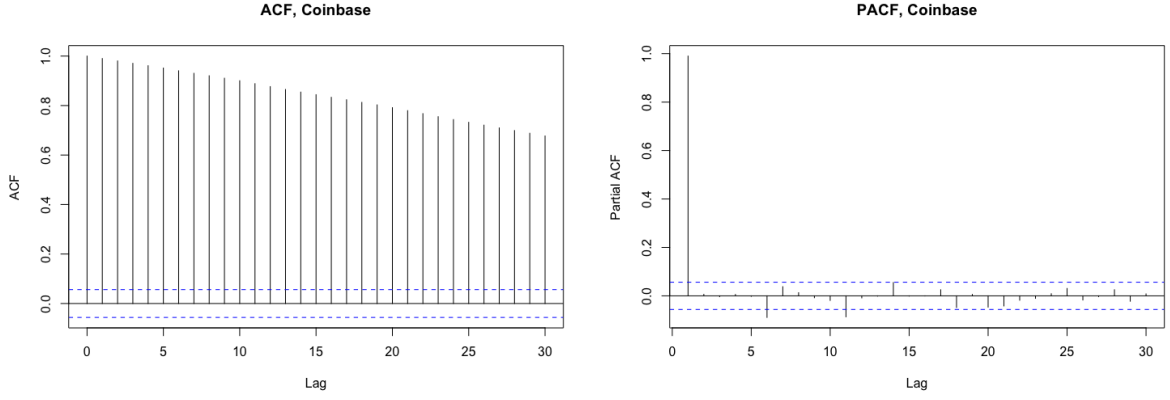
〈표 2〉 Dickey–Fuller 단위근 검정 결과 *p*값

거래소	0차분	1차분	2차분
빗섬	0.05	0.00*	0.00*
코인베이스	0.10	0.00*	0.00*

*: 1%수준에서 유의함



〈그림 3〉 빗썸 가격 데이터의 ACF, PACF(좌, 우)

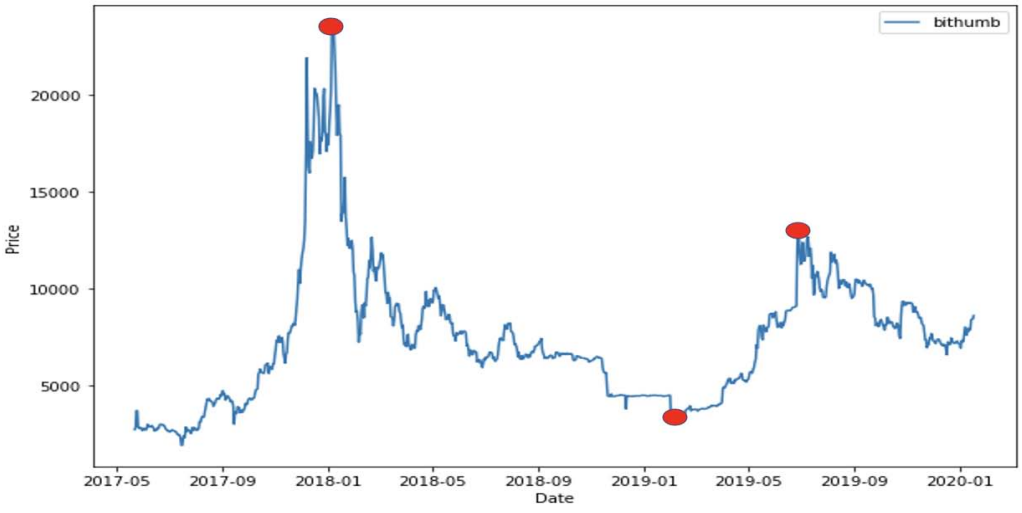


〈그림 4〉 코인베이스 가격 데이터의 ACF, PACF(좌, 우)

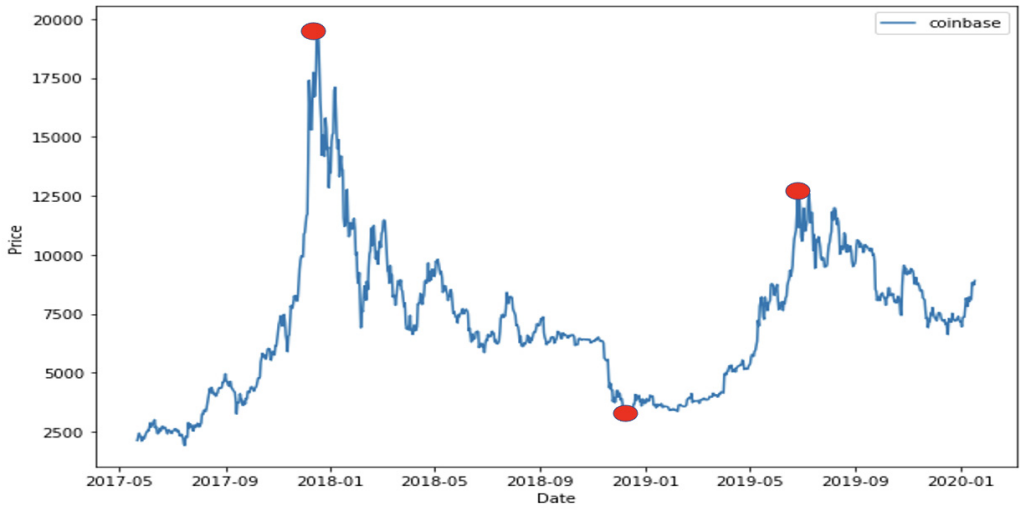
단위근 검정 결과로 두 거래소 데이터 모두 1차분을 할 경우부터 p값이 0.05보다 작기 때문에 d는 1로 결정했다. P와 q는 자기상관함수(Auto Correlation Function, ACF)와 부분 자기상관함수(Partial Auto Correlation Function, PACF)를 통해 결정할 수 있다. <그림 3>은 빗썸 거래소 가격 데이터의 ACF와 PACF이고, <그림 4>는 코인베이스 거래소 가격 데이터의 ACF와 PACF이다. 두 거래소 모두 ACF의 경우 서서히 0으로 감소하는 형태고, PACF의 경우 1차에 두드러지는 스파이크가 나타난 후 모두 0에 까깝게 절단되는 형태를 나타나기 때문에 p는 1, q는 0으로 설정했다.

4.2.2. 순환 신경망 모형

순환 신경망 모형은 Python 3.7.1 환경에서 다차원 신경망 API인 Keras를 사용했다. 데이터의 60%를 모형 학습을 위한 학습데이터로 사용했고, 20%를 하이



〈그림 5〉 빗썸 비트코인 가격 추세 변화 점



〈그림 6〉 코인베이스 비트코인 가격 추세 변화 점

〈표 3〉 첫 번째 추세 변화 점의 용어 기반 키워드 사전 예시

한국어 키워드	영어 키워드
상승	Cryptocurrency
유가	Moon
디지털 화폐	Stocks and Bonds
호재	Bull
뉴욕증권거래소	Securities
페이스북	futures

〈표 4〉 두 번째 추세 변화 점의 용어 기반 키워드 사전 예시

한국어 키워드	영어 키워드
의혹	Tax
지연	Regulation
거래세	Short Selling
규제	Worth
업비트	Track
거래소	Bitcoin

퍼파라미터 튜닝을 위한 검증 데이터, 20%를 성과 평가를 위한 평가데이터로 사용했다. 중요한 모형 모수 중 하나인 Epoch은 1부터 500까지 1씩 늘려가며 최적의 Epoch을 설정했고, 검증 데이터를 활용해 MAE(Mean Absolute Error)를 낮추는 방식으로 진행했다.

4.2.3. 분리 학습 모형(Separated RNN)

먼저 데이터를 나누기 위해 학습데이터에서 추세가 변하는 점들을 설정한다. <그림 5>와 <그림 6>은 본 연구에서 거래소 별로 설정한 추세 변화 점을 표시한 그림이다. 빗썸과 코인베이스 거래소 모두 3개의 점을 추세 변화 점으로 지정을 했다. 이어서 지정한 추세 변화 점들을 기준으로 데이터 셋을 네 부분으로 분리했고, 분리된 데이터 셋 별로 각각 순환 신경망을 적용해 4개의 모형을 구축했다. 본 연구에서는 각각의 모형의 이름을 Separatd_Model_1, Separated_Model_2, Separatd_Model_3, Separatd_Model_4, 전체 모형을 Separated RNN이라 명명한다. 추후 평가데이터에서 유사한 추세를 가지고 있는 데이터에 적용해 예측 성과를 산출한다.

이어 데이터 수집 단계에서 수집한 뉴스데이터를 추세 변화 점 별로 키워드를 정리한다. 본 연구에서는 추세 변화 점과 매칭되는 날짜에 해당하는 뉴스 기사의 제목 및 키워드를 모두 정리해 사용했다. <표 3> 첫 번째 추세 변화 점의 용어 기반 키워드 사전이고, <표 4>는 두 번째 추세 변화 점의 사전이다. 현재 단계에서 구축한 용어 기반 키워드 사전은 추후 예측할 데이터, 즉 평가 데이터의 추세 변화 점 탐색에 활용된다.

4.2.4. 추세 변화 점 탐색 및 예측

마지막 단계에서는 예측할 평가 데이터의 추세 변화 점을 찾기 위해 빈도 기반 텍스트 분석을 사용한다. 먼저 평가 데이터의 일일 뉴스 데이터의 키워드를 키워드의 종류와 종류 별 개수로 정리한다. 이어서 추세 변화 점 별 용어 기반 키워드 사전에 존재하는 키워드의 종류와 종류 별 개수로 정리해 높은 빈도를 가지는 일을 특정 추세 변화 점으로 설정한 후 데이터를 나누게 된다. 이어서 특정 추세 변화 점과 매칭되는

〈표 5〉 첫 번째 추세 변화 점의 키워드가 평가 데이터 일별 뉴스에 존재하는 빈도(한국)

날짜	빈도
2020-01-20	3051
2020-01-21	2807
…	…
2020-02-14	**4142**
…	…
2020-09-16	2611

〈표 6〉 첫 번째 추세 변화 점의 키워드가 평가 데이터 일별 뉴스에 존재하는 빈도(미국)

날짜	빈도
2020-01-20	291
2020-01-21	591
…	…
2020-02-15	**1124**
…	…
2020-09-16	426

Separated_Model_N 모형을 적용해 예측 값을 산출한다. <표 5>와 <표 6>은 한국과 미국 뉴스 기사에서 평가 데이터의 첫 번째 추세 변화 점의 키워드의 일별 빈도 분포를 나타낸다. 한국 데이터의 경우 2020년 2월 14일의 키워드 빈도가 가장 높아 2020년 2월 14일을 기준으로 데이터 셋을 분리하고 이에 맞는 모형을 적용하게 된다. 미국 데이터의 경우 2020년 2월 13일의 키워드 빈도가 가장 높아 2020년 2월 13일을 기준으로 데이터 셋을 분리한다. 이처럼 한국과 미국 뉴스 기사 간에 차이가 존재 했지만, 비슷한 시기의 추세 변화 점을 찾았다.

5. 실험 결과

5.1. 예측 성과 지표

본 실험에서 사용되는 예측 성과 지표는 MAE(Mean Absolute Error), MAPE(Mean Absolute Percentage Error), RMSE(Root Mean Squared Error)이다. 먼저 MAE의 경우 오차 값에 절대값을 적용하기 때문에 가장 직관적으로 알 수 있고, MSE(Mean Squared Error)에 비해 특이 값에 강건하다(Hyndman & Koehler, 2006). 하지만 절대값을 적용하기 때문에 모델이 과소평가된 예측 값을 나타내는지 과대평가된 예측 값을 나타내는지 알 수 없다. MAPE의 경우 MAE와 마찬가지로 MSE보다 특이 값에 강건하지만 0 근처의 값에서는 사용하기 어렵다는 단점이 있다(Hyndman & Koehler, 2006). 마지막으로 RMSE는 오류 지표를 실제 값과 유사한 단위로 다시 반환하기 때문에 해석이 쉬워진다는 특징을 가지고있다. 본 연구에서는 성과 지표에 따른 해석의 편향을 방지하기 위해 MAE, MAPE, RMSE를 모두 계산하고 종합적으로 성과를 평가한다.

5.2. 예측 성과

<표 7>은 ARIMA, 순환 신경망 모형과 Separated_RNN 모형의 예측 성과를 나타낸다. 빗썸 거래소와 코인베이스

〈표 7〉 모형 별 예측 성과

거래소	모형	RMSE	MAE	MAPE
빗썸	ARIMA	1474.62	1114.64	16.86
	순환 신경망	1030.97	810.85	8.89
	Separated_RNN	894.79	715.28	8.64
코인베이스	ARIMA	1358.56	1008.31	15.21
	순환 신경망	842.25	663.75	7.17
	Separated_RNN	717.12	585.77	7.04

〈표 8〉 순환 신경망, 분리 학습 모형의 대응 표본 T검정 결과

거래소	대응 표본	T-통계량	RMSE	MAE	MAPE
빗썸	순환 신경망& Separated_RNN	평균차이	136.18	95.57	0.25
		유의확률	0.00*	0.00*	0.00*
		표본개수	242	242	242
코인베이스	순환 신경망& Separated_RNN	평균차이	125.13	77.98	0.13
		유의확률	0.66	0.64	0.13
		표본개수	242	242	242

*: 1%수준에서 유의함

거래소 모두에서 ARIMA < 순환 신경망 < Separated_RNN 모형 순으로 예측 성과가 우수했으며, 코인베이스 거래소 가격 데이터의 예측 성과가 빗썸 거래소 데이터의 예측 성과보다 세 모형 모두에서 우수했다.

<표 8>은 거래소 별 평가 지표 별 대응 표본 t-test 결과이다. 유의 확률을 살펴보면, 빗썸 거래소의 경우 순환 신경망 모형과 분리 학습 모형 간의 예측 성과에 유의미한 차이가 존재하는 것을 알 수 있으나, 코인베이스 거래소의 경우 통계적으로 유의미한 차이를 보이지 않음을 확인할 수 있다. 이는 코인베이스 거래소의 경우 뉴욕타임스 API를 활용한 미국 뉴스 키워드 추출의 단어가 한국 뉴스 데이터의 키워드 개수보다 양과 질적인 측면에서 차이가 존재했고, 이는 추세 변환 점을 탐색할 때 충분한 정보를 제공하지 않았음을 의미할 수 있다.

6. 결론

본 연구는 ARIMA, 순환 신경망 등 시계열 예측 기법을 활용한 비트코인 가격 예측 모형이 갖는 한계를 제시하고, 한계를 보완하기 위해 학습기간을 분리한 후 복수의 모형을 만들어 국가별로 예측에 활용하는 분리 학습 모형을 제시하였다. ARIMA, 순환 신경망 등 시계열 예측 기법이 갖는 한계는 장기간의 데이터를 학습데이터로 사용할 경우 가격의 추세가 변하는 점이 다수 존재할 수 있고, 추세 변화 점이 학습 및 예측 성과에 악영향을 줄 수 있다는 점이었다. 분리 학습 모형의 성능을 전체 데이터 셋을 사용해 학습한 모형들과 비교한 결과 한국 거래소의 데이터와 미국 거래소의 데이터 모두에서 가장 높은 예측 정확도를 보였다. 이러한 연구 결과를 요약하면 다음과 같다.

첫째, 분리 학습 아이디어에 기반한 분리 학습 모형을 통하여 모형의 예측 성과를 향상시킬 수 있다. 추

세가 변하는 점인 추세 변화 점을 지정하고, 지정한 점들을 기준으로 기간을 분리해 각각 학습시키면 특정 추세를 잘 반영하는 모형들을 구축할 수 있었다. 가격의 상승, 하락, 횡보 등을 반영하는 각각의 모형들은 전체 데이터를 사용해 학습한 모형보다 추세를 더 잘 반영했고, 예측 성과 향상에도 공헌할 수 있다. Azari(2019)는 비트코인 가격 예측에 있어 ARIMA 기법을 전체 데이터 셋에 적용한 결과 장기간의 예측에 있어서는 ARIMA 기법이 유효성을 가지지 않는다고 주장했는데, 분리 학습 모형을 사용한다면 시계열 기법들의 한계를 극복할 수 있을 것이라 기대된다.

둘째, 뉴스 데이터를 활용해 추세 변화 점을 효과적으로 탐색할 수 있다. 추세를 잘 반영하는 복수의 모형이 존재하더라도, 예측할 데이터 셋에 추세가 다른 데이터를 이용해 학습한 모형을 적용하게 된다면 예측 성과가 떨어질 수 있기 때문에, 예측할 데이터에 적절한 모형을 적용하는 것이 중요하다 볼 수 있다. 본 연구에서는 학습 데이터의 추세 변화 점 별 키워드 사전을 구축했고, 평가 데이터 기간에 일 별 뉴스 데이터의 키워드가 각 추세 변화 점의 키워드 사전에 존재하는 단어들의 빈도를 계산해 추세 변화 점을 탐색했다. 탐색한 추세 변화 점과 매칭되는 모형을 적용함으로써 비트코인 가격의 예측 성과를 향상시킬 수 있었다.

셋째, 한국과 미국의 가격 데이터 및 뉴스 데이터를 사용해 각각 별개의 모형을 구축함으로써 예측 성과를 비교하고, 추세 변화 점 탐색에 있어 뉴스 데이터의 효과성을 비교했다. 예측 결과 코인베이스의 비트코인 가격을 예측한 모형의 성과가 빗썸의 비트코인 가격을 예측한 모형의 성과보다 더 높았는데, 같은 구조를 가지는 순환 신경망 기법을 적용을 했음에도 예측 성과에 차이가 존재하는 것은 추세 변화 점 탐색의 유효성과 관련이 있을 수 있다. 대응 표본 T-test의 결과로 미루어 볼 때, 미국 비트코인 가격 데이터의 추세 변화 점 탐색에 활용한 뉴스 데이터 및 탐색 기법은 유효성을 가진다고 판단할 수 있다. 하지만 한국 데이터의 대응 표본 T-test 결과는 분리 학습을 적용한 모형의 예측 성과 향상이 통계적으로 유의하지 않는다고 볼 수 있고, 따라서 한국 데이터에 있어서는 추세 변화 점 탐색이 유효성을 가지기 힘들다고 판단할 수 있다.

본 연구가 갖는 한계점 및 향후 연구방향은 다음과 같다.

첫째, 본 연구에서는 대표적인 시계열 기법인 ARIMA기법과 순환 신경망(RNN)기법을 사용해 가격을 예측하는 모형을 구축했다. 하지만 선행 연구에 다수 존재하는 CNN, GRU, LSTM 등 시계열 예측에 더 효과적일 수 있는 기법들이 존재한다. 추후 연구에서는 분리 학습에 시계열 예측 기법을 사용할 때 더 다양한 종류의 기법을 적용해 볼 필요가 있다.

둘째, 뉴스 데이터를 사용해 추세 변화 점을 탐색할 때, 뉴스 기사의 제목과 키워드뿐 아니라 뉴스 본문을 활용하는 등 더 방대한 양의 뉴스 데이터를 사용할 필요성이 있다. 특히 미국 뉴스 데이터의 경우 키워드의 양, 질적 측면에서 한국 뉴스 기사에 비해 우월하다 볼 수 없었고 또한 한국 데이터의 경우 뉴스 본문을 가지고 있었음에도 불구하고 비교를 위해 사용하지 않았다. 추후 연구에서는 뉴스 본문 혹은 더 다양한 신문사의 데이터베이스를 활용하는 등 방대한 텍스트 데이터를 사용할 필요가 있다.

셋째, 추세 변화 점 탐색에 있어 단순 빈도에 기반한 방법이 아닌 뉴스나 트위터 데이터의 감성 분석이나 딥러닝 기반의 텍스트 분석 등 더 다양한 텍스트 마이닝 기법을 활용할 필요가 있다. 본 연구에서의 예측 성과 향상에 추세 변화 점 탐색이 결정적인 역할을 할 수 있기 때문에, 여러 텍스트 마이닝 기법 간 비교가 필요하다.

〈참고문헌〉

[국내 문헌]

1. 권혁준, 김협, 최재원 (2018). 개인 의료정보 보호를 위한 블록체인 적용 방안: 프라이빗 블록 스킴을 중심으로. **지식경영연구, 19**(4), 119-131.
2. 김선웅 (2021). 암호화폐 트레이딩시스템의 수익성 분석. **한국디지털콘텐츠학회 논문지, 22**(3), 555-562.
3. 김효상 (2019). 암호화자산이 국경 간 자본흐름에 미치는 영향. **국제금융연구, 9**(2), 67-90.
4. 배성완, 유정석 (2018). 머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측. **주택연구, 26**(1), 107-133.
5. 정윤경, 하예영, 이혜인, 양희동 (2020). 공유경제 체제로서 컨소시엄 블록체인을 활용한 와인투자 주식플랫폼 프레임워크. **지식경영연구, 21**(3), 45-65.
6. 조보근, 박경배, 하성호 (2020). 기계학습 알고리즘을 활용한 지역 별 아파트 실거래가격지수 예측모델 비교: LIME 해석력 검증. **정보시스템연구, 29**(3), 119-144.
7. 최재원 (2018). 건강정보에 대한 블록체인 기술 응용: 블록체인 기술은 글로벌 건강 정보 이슈에 대해 만병 통치약이 될 수 있는가? **지식경영연구, 19**(4), 187-201.
8. 허인 (2019). 비트코인 시장의 변동성과 전통 금융시장과의 관계. **시장경제연구, 48**(2), 53-87.

[국외 문헌]

9. Abad, C., Thore, S. A., & Laffarga, J. (2004). Fundamental analysis of stocks by two-stage DEA. *Managerial and Decision Economics, 25*(5), 231-241.
10. Azari, A. (2019). *Bitcoin price prediction: An ARIMA approach*. arXiv preprint arXiv:1904.05315.
11. Brandvold, M., Molnár, P., Vagstad, K., & Valstad, O. C. A. (2015). Price discovery on Bitcoin exchanges. *Journal of International Financial Markets, Institutions and Money, 36*, 18-35.
12. Chang, P. C. (2012). A novel model by evolving partially connected neural network for stock price trend forecasting. *Expert Systems with Applications, 39*(1), 611-620.
13. Cheah, E. T., & Fry, J. (2015). Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin. *Economics Letters, 130*, 32-36.
14. Chen, W., Xu, H., Jia, L., & Gao, Y. (2021). Machine learning model for Bitcoin exchange rate prediction using economic and technology determinants. *International Journal of Forecasting, 37*(1), 28-43.
15. Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics, 365*, 112395.
16. Ciaian, P., & Rajcaniova, M. (2018). Virtual relationships: Short-and long-run evidence from BitCoin and altcoin markets. *Journal of International Financial Markets, Institutions and Money, 52*, 173-195.
17. Dillon, M. (1983). *Introduction to modern information retrieval: G. Salton and M. McGill*. McGraw-Hill, New York.
18. Fauzi, M. A., Paiman, N., & Othman, Z. (2020). Bitcoin and cryptocurrency: Challenges, opportunities and future works. *The Journal of Asian Finance, Economics and Business(JAFEB), 7*(8), 695-704.
19. Felizardo, L., Oliveira, R., Del-Moral-Hernandez, E., & Cozman, F. (2019, October). Comparative study of Bitcoin price prediction using WaveNets, Recurrent Neural Networks and other Machine Learning Methods. *In 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), IEEE*, 1-6.
20. Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D., & Giaglis, G. M. (2015). *Using time-series and sentiment analysis to detect the determinants of bitcoin prices*. Available at SSRN 2607167.
21. Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*(3), 685-697.
22. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679-688.

23. Joo, T. W., & Kim, S. B. (2015). Time series forecasting based on wavelet filtering. *Expert Systems with Applications, 42*(8), 3868-3874.

24. Karalevicius, V., Degrande, N., & De Weerdt, J. (2018). Using sentiment analysis to predict interday Bitcoin price movements. *The Journal of Risk Finance, 19*(1), 56-75.

25. Kodama, O., Pichl, L., & Kaizoji, T. (2017, September). Regime change and trend prediction for Bitcoin time series data. *In CBU International Conference Proceedings, 5*, 384-388.

26. Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems, 104*, 38-48.

27. Lamothe-Fernández, P., Alaminos, D., Lamothe-López, P., & Fernández-Gámez, M. A. (2020). Deep learning methods for modeling Bitcoin price. *Mathematics, 8*(8), 1245.

28. Li, Y., Zheng, Z., & Dai, H. N. (2020). Enhancing Bitcoin price fluctuation prediction using attentive LSTM and embedding network. *Applied Sciences, 10*(14), 4872.

29. Lu, N., Lin, H., Lu, J., & Zhang, G. (2012). A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics, 10*(2), 1659-1665.

30. Luther, W. J. (2016). Bitcoin and the future of digital payments. *The Independent Review, 20*(3), 397-404.

31. Makridakis, S. (1994). Time series prediction: Forecasting the future and understanding the past. *International Journal of Forecasting, 10*(3), 463-466.

32. Mittermayer, M. A. (2004, January). Forecasting intraday stock price trends with text mining techniques. *In 37th Annual Hawaii International Conference on System Sciences, IEEE*, 10.

33. Pawar, K., Jalem, R. S., & Tiwari, V. (2019). Stock market price prediction using LSTM RNN. In *Emerging trends in expert applications and security* (pp. 493-503). Springer, Singapore.

34. Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017, September). Stock price prediction using LSTM, RNN and CNN-sliding window model. *In 2017 International Conference on Advances in Computing, Communications and Informatics(ICACCI), IEEE*, 1643-1647.

35. Tang, X., Yang, C., & Zhou, J. (2009, September). Stock price forecasting by combining news mining and time series analysis. *In 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, IEEE, 1*, 279-282.

36. Telli, Ş., & Chen, H. (2020). Structural breaks and trend awareness-based interaction in crypto markets. *Physica A: Statistical Mechanics and Its Applications, 558*, 124913.

37. Velankar, S., Valecha, S., & Maji, S. (2018, February). Bitcoin price prediction using machine learning. *In 2018 20th International Conference on Advanced Communication Technology(ICACT), IEEE*, 144-147.

38. Wolla, S. A. (2018). *Bitcoin: Money or financial investment?* Page One Economics®.

39. Yao, W., Xu, K., & Li, Q. (2019, June). Exploring the influence of news articles on Bitcoin price with machine learning. *In 2019 IEEE Symposium on Computers and Communications(ISCC), IEEE*, 1-6.

저 자 소 개

원 종 관 (Jonggwan Won)

부산대학교 경영학과에서 학사학위를 취득하였다. 현재 부산대학교 경영학과 경영정보전공 석사과정에 재학 중이다. 주요 관심분야는 지능형 테크핀, 암호화폐 예측, 딥러닝, AI 등이다.

홍 태 호 (Taeho Hong)

현재 부산대학교 경영학과 교수로 재직 중이다. KAIST에서 경영정보시스템을 전공하여 공학석사와 공학박사를 취득하였다. 주요 관심분야는 비즈니스 애널리틱스, 딥러닝, 오피니언 마이닝, CRM 등이다. 주요 논문을 Expert Systems, Expert Systems with Applications, Information Processing & Management, Asia Pacific Journal of Information Systems, 정보시스템연구 등에 게재하였다.

〈 Abstract 〉

The Prediction of Cryptocurrency on Using Text Mining and Deep Learning Techniques : Comparison of Korean and USA Market

Jonggwan Won*, Taeho Hong**

In this study, we predicted the bitcoin prices of Bithum and Coinbase, a leading exchange in Korea and USA, using ARIMA and Recurrent Neural Networks(RNNs). And we used news articles from each country to suggest a separated RNN model. The suggested model identifies the datasets based on the changing trend of prices in the training data, and then applies time series prediction technique(RNNs) to create multiple models. Then we used daily news data to create a term-based dictionary for each trend change point. We explored trend change points in the test data using the daily news keyword data of testset and term-based dictionary, and apply a matching model to produce prediction results. With this approach we obtained higher accuracy than the model which predicted price by applying just time series prediction technique. This study presents that the limitations of the time series prediction techniques could be overcome by exploring trend change points using news data and various time series prediction techniques with text mining techniques could be applied to improve the performance of the model in the further research.

Key Words: Cryptocurrency, Frequency based Text Analysis, Recurrent Neural Networks, Separate Learning, Time series analysis

* Pusan National University

** Pusan National University