JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 5, pp. 47-56, May 2020
https://doi.org/10.9708/jksci.2020.25.05.047

Stock prediction using combination of BERT sentiment Analysis and Macro economy index

Euna Jang*, HoeRyeon Choi*, HongChul Lee*
*Master, Dept. of Industrial Engineering, Korea University, Seoul, Korea
*Visiting Professor, Dept. of Industrial Engineering, Korea University, Seoul, Korea
*Professor, Dept. of Industrial Engineering, Korea University, Seoul, Korea

[Abstract]

The stock index is used not only as an economic indicator for a country, but also as an indicator for investment judgment, which is why research into predicting the stock index is ongoing. The task of predicting the stock price index involves technical, basic, and psychological factors, and it is also necessary to consider complex factors for prediction accuracy. Therefore, it is necessary to study the model for predicting the stock price index by selecting and reflecting technical and auxiliary factors that affect the fluctuation of the stock price according to the stock price. Most of the existing studies related to this are forecasting studies that use news information or macroeconomic indicators that create market fluctuations, or reflect only a few combinations of indicators. In this paper, this we propose to present an effective combination of the news information sentiment analysis and various macroeconomic indicators in order to predict the US Dow Jones Index. After Crawling more than 93,000 business news from the New York Times for two years, the sentiment results analyzed using the latest natural language processing techniques BERT and NLTK, along with five macroeconomic indicators, gold prices, oil prices, and five foreign exchange rates affecting the US economy Combination was applied to the prediction algorithm LSTM, which is known to be the most suitable for combining numeric and text information. As a result of experimenting with various combinations, the combination of DJI, NLTK, BERT, OIL, GOLD, and EURUSD in the DJI index prediction yielded the smallest MSE value.

▸ **Key words**: Natural Language, Sentiment Analysis, Stock Prediction, Macro Economy Index, Deep learning

[요 약]

주가지수는 한 국가의 경제 지표뿐만 아니라 투자판단의 지표로도 활용되므로 이를 예측하는 연구가 지속해서 진행되고 있다. 주가지수 예측을 하는 작업은 기술적, 경제적 및 심리적 요인 등이 반영된 것으로 예측의 정확도를 위해서는 복합적 요인을 고려해야 한다. 따라서 지수의 변동에 영향을 미치는 요인들을 선별하여 반영한 주가지수 예측모델연구가 필요하다. 이와 관련한 기존 연구에서는 시장의 변동을 만들어 내는 뉴스 정보 또는 거시 경제 지표를 각각 이용하거나, 몇 가지의 지표 조합만을 반영한 예측 연구가 대부분이었다. 따라서 본 연구에서는 미국 다우존스지수 예측을 위해 뉴스 정보의 감성 분석과 다양한 거시경제지표를 고려하여 효과적인 지표 조합을 제시하고자 한다. 뉴스 정보의 감성 분석은 최신 자연어처리 기법인 BERT와 NLTK VADER를 사용하고, 예측모델은 주가예측모델로 적합하다고 알려진 딥러닝 예측모델 LSTM을 적용하여 가장 효과적인 지표 조합을 제시했다.

▸ **Key words**: 자연어처리, 감성분석, 주가예측, 거시경제지표, 딥러닝

• First Author: Euna Jang, Corresponding Author: Euna Jang
*Euna Jang (sophie.euna.jang@gmail.com), Dept. of Industrial Engineering, Korea University
*HoeRyeon Choi (kdhong@korea.ac.kr), Dept. of Industrial Engineering, Korea University
*HongChul Lee (hclee@korea.ac.kr), Dept. of Industrial Engineering, Korea University
• Received: 2020. 05. 04, Revised: 2020. 05. 26, Accepted: 2020. 05. 26.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

주식 시장에서 가격이 형성될 때, 가격은 시장에 존재하는 모든 정보를 이미 내포하고 있다. 주식은 기술적 지표를 비롯하여 다양한 거시경제 지표에 반응하기도 하지만 회사의 정보를 포함하는 뉴스 데이터도 영향을 받는 것으로 알려져 있다. 다양한 형태의 데이터는 많은 투자자에게 노출되며 영향력 있는 정보를 추출하기 위한 주가 예측 연구는 꾸준히 이루어지고 있다[1-2].

주가 예측을 위한 기존 연구는 기술적 지표, 거시경제 지표 및 뉴스 데이터를 각각 적용한 연구가 많았으며, 대다수 연구는 각 지표 중에서 주가지수에 영향력이 높은 지표를 선정하여 예측모델에 적용하였다. 특히 기술적 지표를 사용한 연구가 많았다[3]. 주가 방향 예측을 위해 기술적 지표를 입력으로 단일분류기와 앙상블 분류기를 비교한 논문이 있으며[4], 기술적 지표와 뉴스 정보를 결합하여 예측모델의 정확도를 높이고자 하는 연구가 있다[2][5-6]. 그에 비해 거시경제 지표와 뉴스 등의 사회적, 심리적 정보를 조합한 연구는 미비한 편이다. 거시경제 지표를 이용한 논문으로 [7]은 거시경제 지표 중 주가지수와 관련 높은 지표를 선정하고 선택된 지표들이 미국 대표 주가지수인 S&P500과 다우존스지수(Dow Jones Index; DJI)에 끼치는 효과와 영향을 연구했다. 이 연구에서 선택된 거시경제 변수인 생산자 물가 지수, 산업 생산 지수와 유가는 S&P500과 DJI에 강력한 영향력이 있음을 입증했다. [8]은 금융 외환 시장의 불확실성을 나타내는 환율, 금리, 주가지수 등의 연관성에 관한 시계열 모형을 수립하고 Quasi 우도추정법을 이용하여 모수 추정을 시행했다. [9]는 유가 인상이 유럽 주식 수익률에 부정적인 영향을 미침을 보여주었다. [10-11]에서는 유가와는 별도로 유가 변동성이 높을수록 주식 수익률에 부정적인 영향을 미친다는 결과를 보여주었다. [12]는 오일쇼크(oil shock)가 전 세계 주식 시장의 상태 예측에 지대한 영향력을 행사하고 주식의 하향세가 보임을 밝혔다.

또한, [13]은 금융위기나 COVID-19 같은 국제적 쇼크가 발생할 때, 주가와 국제 금값은 음의 상관관계가 형성됨을 입증하였다. 주가와 외환 상관관계 또한 오랫동안 연구되었다. 외환 시장의 변동은 국제적 수출입의 경쟁력에 영향을 미치며 주가와 밀접한 관계가 있다고 한다. 중국이 2011년 World trade organization(WTO)에 가입한 이후 미국의 경제 변수에 많은 영향을 미친다고 하였는데[14], 이러한 발견은 투자 관점에서 경제적으로 중요한 요인이다. 또한, 뉴스 정보는 G-7 (USA, Canada, Japan, German, UK, Italy, France)나라들의 환율과 주가지수와의 상관관계에서 단기적으로 민감하게 반응하며, 이러한 결과는 위험 관리, 정책 결정, 국제투자자에게 큰 도움이 될 것이다[15]. 미국과 신흥국 4곳 (중국, 인도, 브라질, 멕시코)의 환율 관계를 연구한 결과, 중국과 인도의 환율이 미국에 가장 큰 연관성이 있음을 보였다[16]. [17]은 금융위기 상황에서 유럽연합(EU)과 미국의 환율이 주식 시장에 미치는 상호영향력 관계를 연구하였다. 결과적으로, EU와 미국의 장단기 환율이 주가지수와 인과 관계가 있음을 밝혔다. 이 모든 연구는 국제 금값을 비롯하여 유가와 5가지 환율(중국, 일본, 인도, 캐나다, 유로)이 주가 분석 및 예측에 영향력 있는 요소임을 입증하였다.

인공지능 기술의 발달로 온라인 뉴스와 소셜 미디어 정보 이용 및 분석이 수월해짐에 따라 주가 예측에 뉴스 및 소셜 미디어 정보를 적용하는 연구가 진행되고 있다. [18]은 주가에 영향을 미치는 트위터의 감성을 분석하기 위해 2진으로 분류 가능한 OpinionFinder (OF)와 6가지로 감성 상태(clam, alter, sure, vital, kind, happy)를 나타낼 수 있는 GPOMS를 이용하였다. [1]과 [4]는 주가를 예측하기 위해 뉴스 데이터를 사용하였고 [2]는 주식의 가격 정보와 뉴스 기사의 감성 정보를 이용하여 NASDAQ100 지수에 등재된 20개 회사의 포트폴리오 추세를 예측할 수 있는 예측모델을 개발하였다. [19]는 소셜 미디어 등의 웹 데이터에서 불안, 걱정, 두려움의 감정 표현이 증가함에 따라 S&P500 지수가 하락하는 경향이 있음을 보였다. [20]은 트위터 피드를 통해 분석된 대중의 분위기가 DJI와 밀접한 관련이 있음을 보여주었다. [21]은 금융 뉴스와 시계열 데이터를 통합하여 주식 분석을 시행하여 전 주가를 예측하는 연구를 수행하였다. 이 모든 연구는 뉴스와 소셜 미디어 정보가 주가 분석 및 예측에 영향력 있는 요소임을 입증하였다.

본 논문에서 지수예측의 대상이 되는 DJI는 효율적 시장 가설 중 약형 효율적 시장 가설에 속하는 것으로 알려져 있다. 약형 효율적 시장 가설은 현재의 시장에서 거래가 가능한 주식, 채권 및 유형 자산 등의 가격에 이용 가능한 모든 과거 주가에 대한 정보가 이미 반영된 것을 말한다. 즉, DJI는 과거의 주가 정보를 담고 있는 기술적 분석 이외의 외부 정보를 활용하여 주가 분석을 하는 것이 효과적이다[24].

따라서 본 연구에서는 미국 DJI 예측의 정확도 향상을 위해 뉴스 정보의 감성 분석과 다양한 거시경제지표의 효과적 조합을 딥러닝 예측모델을 통해 제시하고자 한다. 주가지수 예측에 사용한 모델은 시계열 데이터와 비정형 데이터인 뉴스 정보를 결합한 데이터를 이용하기에 적합

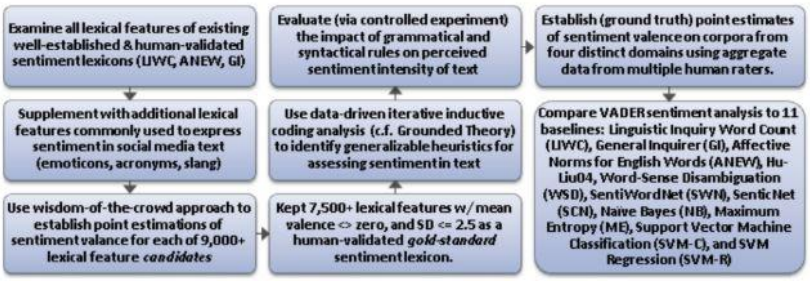

Fig. 1. VADER Methods and Process approach Overview[22]

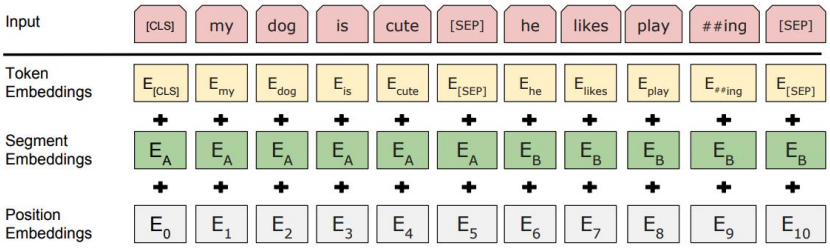

Fig. 2. BERT Input representation[23]

하다고 알려진 딥러닝 모델인 LSTM을 사용하였다.

본 논문과 기존 논문들과의 차이점은 다음과 같다. 첫 번째, DJI에 영향력이 큰 최신의 거시경제지표를 선별하였다. 기존 논문들은 2000년대와 2010년의 초반에 발표한 연구와 자료를 기반으로 지표를 선정하였다. 반면에 본 논문은 과거의 자료뿐만 아니라 최근의 연구와 기관 및 미디어에서 발표한 자료를 기초로 거시경제지표를 선별하였다. 예를 들어 기존 연구 당시에는 제외되었던 인도와 중국의 환율을 새로운 지표로 선정하였다. 또한, 현재까지도 미국경제에 영향력을 행사하는 일본, 캐나다 및 유럽연합의 환율과 국제 금값 및 유가를 거시경제지표로 삼았다. 두 번째는 자연어처리 분야에서 기존 연구에 사용했던 Bag-of-Words나 Word2Vec 모델보다 우수한 성능을 내는 최신 모델로 알려진 BERT를 사용하였다. 대다수 논문은 한 가지 모델을 사용하여 뉴스의 감성 또는 문맥을 분석하였으나, 본 연구는 규칙 및 어휘 기반의 감성 분석 방법으로 알려진 NLTK VADER를 BERT와 함께 지표로 사용하여 두 모델의 상호보완 효과를 예측에 반영하였다.

본 논문의 2장은 연구에 사용된 감성 분석 기법과 딥러닝 예측모델을 서술하며, 3장은 연구에 사용된 데이터에 대한 설명을, 4장은 연구모형과 실험 결과를 기술하고, 결론 및 향후 연구는 마지막 장에서 서술하였다.

II. Related Work

1. Stock prediction using Sentiment analysis

감성 분석은 자연어처리 기술 중 하나로, 텍스트에 담긴 사항을 이해하고 극성을 분석하여 정보를 추출하는 방법이다. 감성 분석은 3가지 방법론으로 나눌 수 있다. ① 규칙 및 어휘 기반(Lexicon based), ② 머신러닝 기반(Machine learning), ③ 딥러닝(Deep learning) 기반의 감성 분석이다[28]. 본 논문에서는 규칙 및 어휘 기반의 감성 분석 방법인 NLTK VADER(Valence Aware Dictionary and sentiment Reasoner)와 딥러닝 기반의 BERT 모델 2가지를 사용하여 감성 점수를 부과했다.

먼저 NLTK VADER는 NLTK에 내장된 모듈로, 사전 기반 감성 분석(Lexicon-based sentiment analysis) 도구이다. 사전 기반 방식에서는 긍정 단어와 부정 단어에 대한 점수 리스트가 미리 존재한다. 사전 기반 감성 분석에서는 전처리가 매우 중요하게 여겨지고, 사전을 바탕으로 감성 점수를 산출하기 위해 문장 혹은 문서에 속한 단어를 토대로 해당 문장이나 문서에 대한 점수를 계산한다. VADER의 장점은 크게 이모티콘, 비속어, 축약어, 감정 기호 등에 대한 감정 분석을 할 수 있고, 감정의 강도를 제공하여 상대적 비교가 가능한 것이다[22](Fig. 1).

BERT(Bidirectional Encoder Representations for Transformers) 모델은 2018년 Google에서 개발된 딥러닝 모델로 최근 다양한 자연어처리 분야에서 가장 우수한

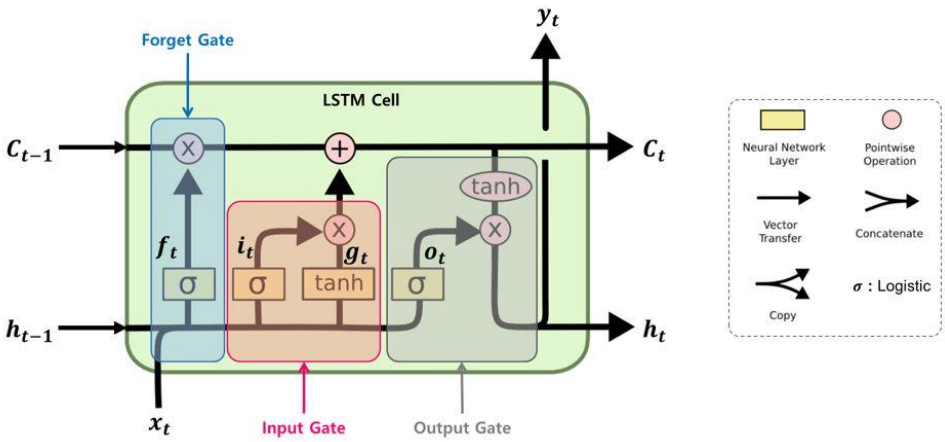

Fig. 3. LSTM model [31]

성능을 보였다. BERT는 트렌스포머(Transformer)에 기반을 둔 모델로 사전학습 후 특정 목적을 위해 Fine-Tuning하여 여러 가지 자연어처리 문제에 적용될 수 있다. 사전학습은 레이블이 없는 코퍼스를 이용하여 비지도 방식으로 진행되고, Fine-tuning을 통해 목적에 맞게 학습하는 것이 특징이다. 이전 연구와는 다르게 양방향 모델이 문장의 앞뒤 문맥을 고려하여 더 높은 정확도를 보인다. 파라미터의 수에 따라 Base와 Large 2개 모델로 나누어지며, 본 논문은 Base 모델을 이용하였다. Fig. 2처럼 BERT의 입력은 세 가지의 임베딩(Embedding)으로 표현된다. 첫째, 토큰 임베딩(Token Embeddings)은 토큰의 의미 표현을 나타내며, 기존에 많이 사용된 워드임베딩 방식이 아닌 Word Piece 임베딩 방식을 사용한다. 이 방식은 기존에 존재하던 Out of Vocabulary (OOV) 처리에 효과적이며 정확도 상승효과를 보여준다. 둘째, Segment Embeddings은 문장과 문장을 문장 구분자([SEP])으로 이어주는 용도로 표현한다. 2개의 문장을 문장 구분자로 이어진 하나의 문장(Single Sequence)로 처리하는데, 두 문장을 구분하기 위해 각각 문장을 A와 B의 Embedding이라 한다. 셋째, Position Embeddings은 단어들의 절대적인 위치 정보를 나타낸다. BERT는 대용량의 코퍼스를 사전학습하기 때문에 긴 시간과 비용이 들지만 이미 학습된 모델을 이용하여 시간을 효율적으로 사용할 수 있다[23](Fig. 2).

따라서, 본 논문에서는 NLTK VADER와 BERT의 감성분석을 각각 지표로 사용함으로써 두 모델의 상호보완적 효과를 얻었으며 이는 주가지수 예측의 정확도를 높였다.

2. Stock prediction model

오래전부터 주가 예측 연구는 끊임없이 진행됐다. 하지만 수많은 변수와 정보는 반영하기 쉽지 않다. 주가 분석 방법에는 크게 기술적 분석(Technical Analysis)과 거시적 경제 분석(Macroeconomics Analysis)으로 나눌 수 있다. 기술적 분석은 간단히 말하자면 과거 주가 움직임과 거래량을 바탕으로 예측을 하는 방법이며, 거시적 경제 분석은 주가에 영향을 미치는 변수 즉 금값, 유가, 환율, 물가상승률 등을 고려하여 예측하는 것이 대표적이다.

전통적인 주가 예측 방법론으로는 일변량 자동 회귀(AR), 일변량 이동 평균 (MA), 단순 지수 평활 (SES) 및 ARIMA(Auto regressive Integrated Moving Average) 등을 활용해왔다. 최근 들어 시계열 데이터를 효과적으로 예측하는 신경망 구조기반의 딥러닝 기술들이 주목받고 있다. CNN을 이용하여 다양한 주가 정보를 성공적으로 조합하였다[25]. 일반 신경망 구조기반에 시계열 데이터 개념이 추가된 순환신경망(RNN; Recurrent Neural Network)은 은닉층에 존재하는 과거 정보를 활용한 망이다. 하지만 장기 의존성 문제를 해결하지 못하는 논리가 나오면서 LSTM이 등장했다. LSTM은 RNN과 비슷한 구조로 되어 있지만, 은닉층의 메모리를 얼마나 잊어버릴지 결정할 수 있는 Forget Gate가 존재한다. 이를 이용하여 이전의 정보를 현재의 문제 해결에 활용할 수 있다. 따라서 LSTM을 이용한다면 길고 복잡한 데이터도 손실 없이 학습할 수 있다(Fig. 3).

Table 1. Dow Jones Index

	DJ_Close	DJ_High	DJ_Low	DJ_Open	DJ_Volume
Count	503	503	503	503	503
Mean	25694	25827	25547	25693	320007512
Std	1165	1123	1212	1167	98774147
Min	22445	22879	21713	21858	155970000
25%	24877	24994	24709	24851	250492500
50%	25649	25775	25497	25657	298735000
75%	26447	26535	26322	26438	361730000
Max	28645	28702	28609	28675	900510000

Table 2. Example of Changing News data Date

Headline	Date	Prediction Date
Thanks to a new musical, Cher's favorite designer may finally get the respect he deserves.	2018-12-01T11:00	2018-12-01
The season's scents of fir and pine evoke primal experiences, and remind us what we may lose.	2018-12-01T19:30	2018-12-02

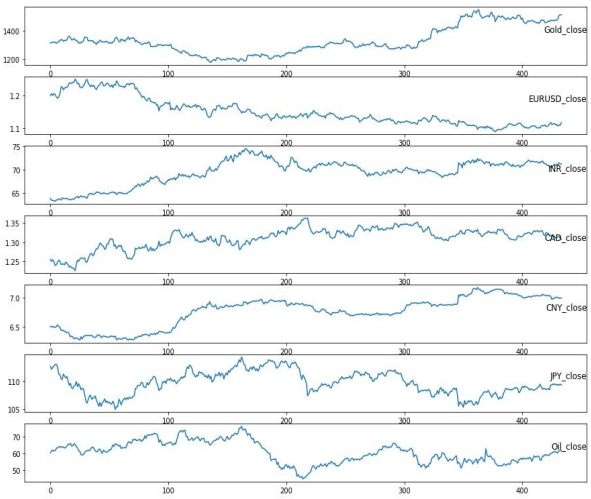

Fig. 4. Macro Variables Close price

기존 연구에 따르면 LSTM은 다른 모델보다 시계열 입력데이터의 특성을 포착하기에 적합하며, 숫자와 텍스트 정보를 결합하기에 적절한 방법론이다[26]. [26]은 LSTM이 변동하는 시계열 변화를 잘 포착할 수 있다고 가정한 후, 이를 이용해 뉴스 벡터와 주가 벡터를 결합하였다. LSTM은 방법론이 복잡한 DCNN보다 경쟁력 있는 구조를 가졌다고 밝혔으며, LSTM와 같은 딥러닝 기반 알고리즘이 ARIMA 모델과 같은 기존 알고리즘보다 우수함을 밝혔다[27-28]. 더 나아가 RNN 및 LSTM (Long Short Term Memory) 네트워크의 구현을 통해 다양한 실험이 성공으로 이루어졌으며, 비선형 딥러닝 방법은 성능이 저조한 ARIMA 보다 예측력이 뛰어나다는 것을 확인했다[29]. 또한, LSTM은 순서를 가진 소리 이미지나 자연어를 처리하는데 효율적이며, 복잡하고 일시적인 특성의 데이터들을 다루는 데 유리하다는 것을 밝혀냈다[30]. 마지막으로, 뉴스 데이터와 시계열 데이터를 통합하기 적절한 알고리즘을 선택하기 위해 LSTM, ANN, SVR 각각의 성능 비교 결과, LSTM 방법론이 가장 좋은 예측 결과를 보였다[24].

본 논문에서는 시계열 분석과 비정형 데이터인 뉴스를 함께 통합하기 위해 순환 신경망 중에서도 가장 많이 쓰이는 LSTM을 방법론으로 채택하여 은닉층에서의 장기의존성(Long Term Dependency) 문제를 효율적으로 처리하고 기존 예측모델의 한계점을 극복하고자 하였다.

III. Data

이번 장에서는 다우존스지수(Dow Jones Index; DJI)의 종가를 예측하기 위해 뉴스 데이터와 거시경제지표 데이터를 어떻게 수집하고 전처리했는지 설명한다.

1. Dow Jones Index

DJI의 종가는 Yahoo Finance를 이용하여 수집하였다. 파이썬의 Pandas Data Reader Module을 통해 get_data_yahoo API를 사용하여 2018년 1월 1일부터 2019년 12월 31일까지의 지수를 수집하였다. 총 503일 동안의 수정종가(Adjusted close), 종가(close), 저가(low), 고가(high), 시가(open), 거래량(volume)을 크롤링 했다(Table 1).

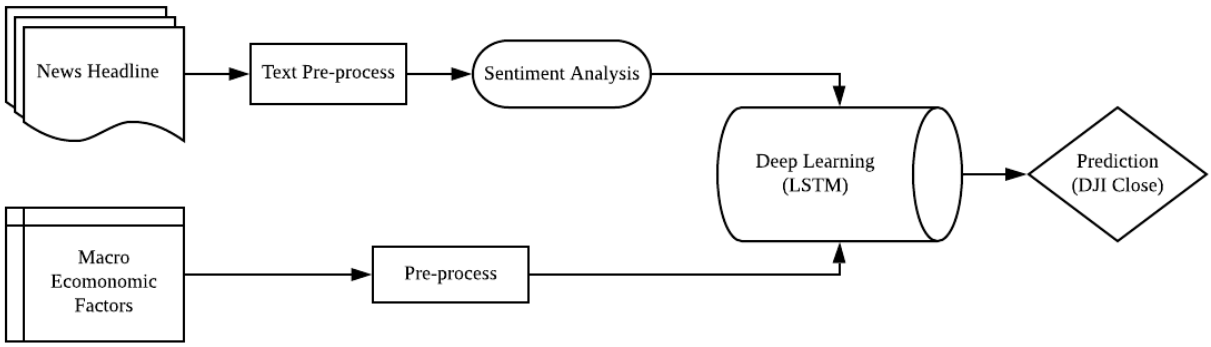

Fig. 5. Model Pipe line

$$DJI_t^{Close\,price} \approx News_t + Gold_t + Oil_t + Ex_t \qquad \text{(Eq. 1)}$$

2. News Data

뉴스 데이터는 New York Times API를 통해 2018년 1월 1일부터 2019년 12월 31일까지 비즈니스(Business) 부문의 신문 표제(Headline) 데이터로 수집되었다. 수집된 데이터는 신문 표제, 날짜 및 시간, 종류, 기사, URL로 이루어진 리스트이며 전처리 후, 총 701일간 93952개의 데이터 중에서 DJI의 예측 기간에 맞추어 공휴일과 토·일요일을 제외한 503일로 정리하였다. 주식 장 마감 시간인 당일 16시부터 다음날의 15시 59분까지 데이터가 다음날의 종가 주가지수에 영향을 미치므로 시간대에 맞게 데이터의 날짜를 바꾸어 주었다(Table 2).

3. Macroeconomic Data

거시경제지표 데이터는 시계열 데이터 형태인 국제 금값, 유가, 환율을 사용하였으며, 이는 Yahoo Finance의 Yahoofinancials Module을 이용하여 일별 자료를 수집하였다. 유가는 Crude Oil (‘CL=F’)을 대표하는 세 가지 원유 중 세계 유가변동의 기준이 되는 미국의 대표적인 원유인 WTI (West Texas Intermediate)를 사용하였다. 환율은 미국경제와 관련이 깊은 5개국의 환율, 5가지 형태를 사용하기 위해 인도, 일본, 중국, 캐나다 및 유럽의 화폐교환비('INR=X', 'JPY=X', 'CNY=X','CAD=X', 'EURUSD=X')를 각각 수집하였다(Fig. 4). DJI와 마찬가지로 503일에 대한 각 환율의 Adjusted close, close, low, high, open price를 수집하였다.

IV. The Proposed Scheme

이번 장에서는 비정형 데이터인 뉴스와 시계열 데이터 다우존스지수(DJI), 금값, 유가와 5개의 환율을 학습모델의 입력값으로 사용하기 위한 통합 방법을 설명한다. 본 논문의 전체 개요는 Fig. 5를 통해 확인할 수 있다.

1. Methods

앞서 언급한 바와 같이 DJI 종가에 영향을 주는 뉴스와 거시경제지표인 국제금값과 유가 및 5가지 환율을 이용하여 DJI 종가를 예측하려 한다. 이를 수식으로 표현하면 (Eq. 1) 같다.

$DJI_t^{Close\,price}$는 DJI의 t 날짜의 close price를 의미한다. $News_t$는 해당 날짜 t에 영향을 주는 모든 뉴스 데이터의 감성 점수 평균을 의미한다, Oil_t 와 $Gold_t$ 는 날짜 t의 Adjusted close, close, low, high, open price를 의미한다. 다양한 유가 중 미국의 가장 대표적인 유가 WTI (West Texas Intermediate)를 사용하였다. Ex_t는 인도/미국, 일본/미국, 중국/미국, 캐나다/미국, 유럽/미국의 환율의 t 날짜에 해당되는 Adjusted close, close, low, high, open price를 의미한다. 환율의 5가지 형태는 다양한 조합을 만들어 변수로 사용되었다. 수집된 모든 거시경제지표는 데이터의 범위를 일정하게 맞추기 위해 0에서 1의 범위 값으로 정규화하였다. 뉴스정보는 NLTK와 BERT 모형을 통해 감성을 추출한다. 이후 이들은 위의 공식처럼 LSTM 알고리즘을 통해 학습하여 통합하였다. 예측 오차는 평균 제곱 오차(MSE;Mean Square Error)와 평균 제곱근 오차(RMSE;Root Mean Square Error)로 측정하였다. MSE는 실제 값과 추정 값의 차이 즉, 잔차(Residuals)가 얼마인가를 알려주는 척도로 사용된다. RMSE는 MSE의 값이 너무 커서 계산 속도가 느려질 때 제곱근을 씌워 시간적 비용을 줄여 준 값이다. 각각의 수식은 (Eq. 2)와 같다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) \qquad \text{(Eq. 7)}$$
$$RMSE = \sqrt[2]{MSE}$$

본 연구에서 목표하는 미국 DJI 예측의 정확도 향상을 위해 뉴스 정보의 감성 분석과 다양한 거시경제지표의 효과적 조합을 딥러닝 기반 예측모델 LSTM을 통해 제시하고자 한다.

2. News Analysis

뉴스 Headline을 NewYorkTimes에서 수집 후, NLTK VADER와 BERT에 사용하기 위해 NLTK 모듈을 이용하여 정규화 및 정제 전처리를 하였다. Poster와 WordNetLemmatizer를 이용한 어간 추출한 후, 대소문자 통합, 불용어(Stopward)를 제거한다. 이때 모든 텍스트는 단어의 원형 형태로 전처리가 되며 감성 분석할 준비가 된다.

본 논문에서 사용한 NLTK VADER 모듈은 문장의 토큰화를 실행한 후 사전을 기반으로 문장 및 문서에 긍정과 부정, 중립 단어의 가중치에 따라 감성 점수를 매기며, 이를 Compounding이라고 표현한다. 본 논문에서는 일반적으로 설정하는 Compounding 범위에 따라 긍정, 부정과 중립을 분류하였다. 즉, Compounding 값이 0.2보다 높으면 긍정인 1, -0.2에서 0.2 사이이면 중립상태인 0, Compounding 값이 -0.2보다 낮으면 부정을 의미하는 -1로 설정하였다. 본 연구에서 수집한 93,952개의 뉴스 기사 중 긍정 31,280개, 중립 36,697개, 부정 25,975개였으며 비율로 환산하면 긍정 33.3%, 중립 39.1%, 부정 27.6%로 집계되었다.

BERT 모델은 대규모 데이터 셋을 사전 학습하여 간단한 Fine-tuning을 통해 여러 가지 자연어 처리 문제를 해결할 수 있도록 고안되었다[23]. 본 연구 과정에서 BERT는 Basic (Transformer black = 12, Hidden layer = 768, Attention head = 12) 옵션을 사용하여 감성 레이블이 있는 데이터로 사전학습하였다. Batch size = 8, Epoch = 100, learning rate = 0.001, dropout rate = 0.3일 때에 정확도(accuracy)가 0.83으로 가장 좋은 결과를 얻었다. 사전학습 된 BERT로 수집한 뉴스 데이터 93,952개에 대한 감성 극성 분류를 진행하였고, 총 93,952개 중 긍정 10,965개, 중립 38,131개, 부정 44,855개의 결과를 얻었다. 각각의 비율은 긍정 11.7% 중립 40.6% 부정 47.7%이다. 각 기사로 생성된 긍정, 중립, 부정은 해당 날짜의 평균 감성 점수로 다시 compounding 하여 학습 데이터로 사용하였다.

감성 분석의 결과값은 DJI 예측 분석의 한 변수로 사용된다. 뉴스 기사의 감성 분석은 BERT, NLTK, BERT+NLTK 3가지 변수 형태를 사용하였다.

3. Prediction model based on Deep learning

DJI의 종가를 예측하기 위해 비정형 데이터와 시계열 데이터를 조합하기에 가장 타당하다고 알려진 딥러닝 방법의 하나인 LSTM 사용한다. 많은 변수를 균등하게 통합하기 위해 각각 변수들을 모두 1차원 형태로 바꾸었으며, NLTK VADER와 BERT의 뉴스 감성, 국제 금값, 유가, 5가지 환율 즉, 9가지 입력값을 조합하여 실험하였다(Fig. 6).

Table 3. Summary of Variable Combination Results

Combination	Loss	Batch Size				
		32	64	128	256	512
Dow Jones (DJ)	MSE	0.001058	0.000846	0.000755	0.000625	0.000564
	RMSE	0.032537	0.029100	0.027493	0.025009	0.023752
DJ + NLTK + BERT	MSE	0.001420	0.001077	0.000676	0.000516	0.000436
	RMSE	0.037693	0.032825	0.026016	0.022715	0.020903
DJ + Gold + Oil + ALL Currency	MSE	0.000913	0.000647	0.000416	0.000256	0.000202
	RMSE	0.030225	0.025442	0.020406	0.016011	0.014227
DJ + NLTK + BERT + Gold + Oil + ALL Currency	MSE	0.000710	0.000479	0.000451	0.000245	0.000187
	RMSE	0.026649	0.021891	0.021251	0.015668	0.013696
DJ + NLTK + BERT + Gold + Oil + EUR	MSE	0.000917	0.000482	0.000550	0.000199	0.000148
	RMSE	0.030288	0.021969	0.023460	0.014112	0.012181
BERT	MSE	0.038951	0.030013	0.016581	0.008398	0.004648
	RMSE	0.197361	0.173243	0.128769	0.091641	0.068181
NLTK	MSE	0.019886	0.017425	0.012136	0.005093	0.002781
	RMSE	0.141018	0.132005	0.110166	0.071371	0.052744

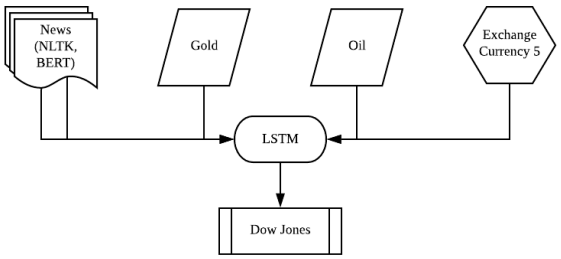

Fig. 6. Deep learning prediction method

각각의 입력값들은 실험 전 정규화를 통해 0에서 1사이 값으로 변형 후 LSTM 알고리즘에 입력하였다. Training과 Test 데이터 셋은 7:3으로 나누고, 각 데이터 셋은 배치 크기(batch size)를 32, 64, 128, 256 및 512로 나누어 학습하였다. LSTM의 입력값은 3차원(size, timestamp, feature)으로 구성되며 size는 입력값의 총 길이, timestamp 는 30, feature는 입력값의 종류를 의미한다, Epoch는 최대 200, Forget gate 즉 dropout은 0.2, 활성화 함수(activation function)는 하이퍼볼릭탄젠트(tanh)를 사용하였다. 예측하고자 하는 Target의 개수가 DJI 종가 하나이므로, Dense에서 출력값을 한 개로 설정하였다. 입력값에 따른 LSTM 모델의 성능을 측정하기 위해서 예측값과 Test 데이터의 오차값을 사용하여 MSE값과 RMSE을 산출하였다. 모든 변수의 조합을 실험하였으며, 영향력 있는 변수의 조합만 비교한 결과는 Table. 3과 같다.

실험 결과, DJI만 지표로 사용했을 경우 NLTK와 BERT를 각각 단독 지표로 사용한 결과 대비 MSE 및 RMSE의 값이 작게 나왔으며(Fig. 7), DJI와 NLTK 및 BERT를 조합한 결과는 DJI만 사용한 결과보다 MSE 값이 0.000128만큼 줄어들었다. 이 결과를 통해 NLTK와 BERT를 함께 사용하는 것이 뉴스 감성 분석에서 서로 협력 작용을 하고 있음을 알 수 있었다. 최종적으로 가장 작은 MSE 값인 0.000148을 얻은 DJI와 국제 금값과 유가 그리고 환율 중 EUR/USD를 포함하여 NLTK VADER와 BERT의 감성 지표가 최적의 조합이었다(Fig. 8)(Table. 3).

V. Conclusions

본 논문은 다우존스지수(DJI) 예측 향상을 위해 딥러닝 모델을 사용하여 거시경제지표와 뉴스 감성 정보와의 효과적인 조합을 제시하였다. 사용한 거시경제지표는 기존 연구에서 입증된 유가, 환율 및 금값을 적용하였으며, 특히 환율은 미국경제에 영향을 미치는 USD/Euro, CAD/USD, CNY/USD, JPY/USD, INR/USD를 사용하였다. 뉴스 정보는 뉴욕 타임즈 경제 관련 헤드라인 기사를 API로 수집하였고, 감성 분석에 사용된 모델은 텍스트 분석에 대표적으로 사용되는 NTLK VADER와 BERT를 사용하여 긍정과 부정 그리고 중립의 세 가지 형태로 각 기사에 대한 감성 분석을 수행하였다. 지표들의 효과적 조합을 찾아내기 위해 시계열 데이터와 비정형 데이터인 뉴스를 함께 통합할 수 있는 딥러닝 알고리즘인 LSTM을 적용하여 실험하였다.

최종적으로 가장 좋은 예측 성능을 보인 지표는 DJI와 국제 금값, 유가 그리고 환율 중 EUR/USD에 NLTK VADER와 BERT의 감성 분석이 모두 사용된 조합이었다. 본 결과를 통해 DJI의 예측에 있어서 NLTK VADER와 BERT는 서로 상호보완적 효과가 있음을 알 수 있었으며, USD/Euro가 다른 환율보다 영향력이 있음을 확인할 수 있었다.

본 논문의 한계점은 첫 번째, 뉴욕 타임즈의 헤드라인 기사만 사용함으로써 기사 전문을 사용한 경우에 비해 감성 분석의 정확도가 떨어질 수 있다는 점과 경제 관련 소셜 미디어의 정보가 제외되어 다각도의 감성 정보가 반영되지 않은 점을 들 수 있다. 두 번째는 DJI라는 도메인에 제한적인 지표 조합을 제시한 것이다. 따라서, 향후 연구는 본 논문의 한계점을 보완하는 연구를 진행하고, 기술적 지표를 추가하고 개별 기업의 주가 예측에 효과적인 지표 조합을 제시하여 트레이딩 알고리즘에 적용할 예정이다.

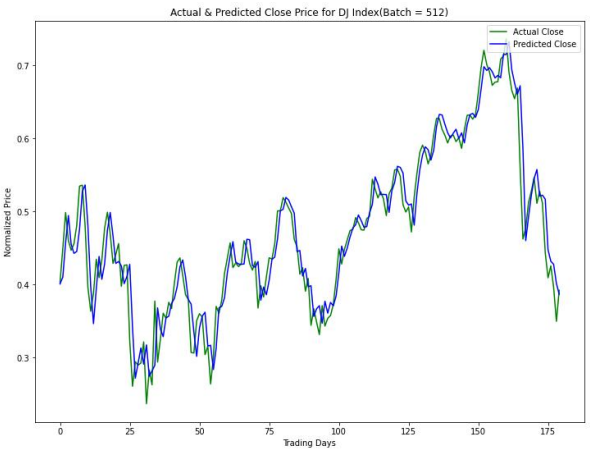

Fig. 7. Dow Jones Index Prediction
(Only DJI)

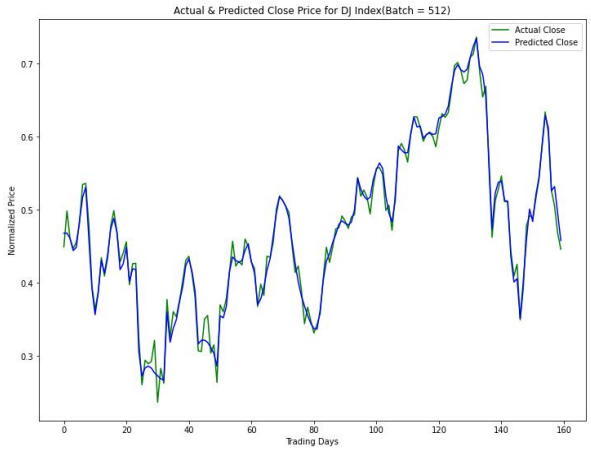

Fig. 8. Dow Jones Index Prediction
(DJI + BERT + NLTK + Gold + Oil + EURUSD)

REFERENCES

[1] Kloptchenko, Antonina, Tomas Eklund, Jonas Karlsson, Barbro Back, Hannu Vanharanta and Ari Visa. “Combining data and text mining techniques for analyzing financial reports,” Int. Syst. in Accounting, Finance and Management, Vol. 12, No. 1, pp. 29-41, March. 2004. DOI: 10.1002/isaf.239

[2] Ratto, Andrea Picasso, Simone Merello, Yukun Ma, Luca Oneto and Erik Cambria. “Technical analysis and sentiment embeddings for market trend prediction,” Expert Syst. Appl. Vol. 135, pp. 60-70, Nov. 2019. DOI: 10.1016/j.eswa.2019.06.014

[3] S. K Kim and H. Y. Kim, “The Study of the Financial Index Prediction Using the Equalized Multi-layer Arithmetic Neural Network,” Journal of The Korea Society of Computer and Information, Vol. 8, No. 33, pp. 113-123, 2013.

[4] Kyun Sun Eo, Kun Chang Lee. “Predicting stock price direction by using data mining methods,” Journal of The Korea Society of Computer and Information, Vol. 22, No. 11, pp. 111-116. Nov. 2017.

[5] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada and A. Sakurai, "Combining Technical Analysis with Sentiment Analysis for Stock Price Prediction," 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, pp. 800-807, Dec. 2011. DOI: 10.1109/dasc2011.138

[6] Zhai. Y. Hsu. A. and Halgamuge. S.K, “Combining news and technical indicators in daily stock price trends prediction,” Advances in Neural Networks Lecture Notes in Computer Science, Vol. 4, pp. 1087–1096, June. 2007. DOI: 10.1007/978-3-540-72395-0_132

[7] Sirucek. Martin. “Macroeconomic variables and stock market: US review,” International Journal of Computer Science and Management Studies. Vol. 12, No. 3, Aug. 2012.

[8] In-Kyu. Kim, “ Study for Exchange rate, Interest, Stock price Using Quasi-Likelihood Estimatorfor,” Journal of The Korea Society of Computer and Information , Vol. 20, No. 1, pp. 255-256, Jan. 2012.

[9] Oberndorfer. Ulrich, “Energy prices, volatility, and the stock market: Evidence from the Eurozone,” Energy Policy, Vol. 37, No. 12, pp. 5787-5795, Dec. 2009. DOI: 10.1016/j.enpol.2009.08.043

[10] Malik. Farooq and Bradley T. Ewing. “Volatility transmission between oil prices and equity sector returns,” International Review of Financial Analysis, Vol 18, No. 3 pp. 95-100, June. 2009. DOI: 10.1016/j.irfa.2009.03.003

[11] Chiou, Jer-Shiou and Yen-Hsien Lee. “Jump dynamics and volatility: Oil and the stock markets,” Energy, Vol. 34, No. 6, pp. 788-796, June. 2009. DOI: 10.1016/j.energy.2009.02.011

[12] Angelidis, Timotheos, Stavros Degiannakis and George N. Filis. “US stock market regimes and oil price shocks,” Global Finance Journal, Vol 28, pp. 132-146, Oct. 2015. DOI: 10.1016/j.gfj.2015.01.006

[13] Smith G, “The price of Gold and Stock Price Indices for the United States,” the World Gold Council, pp. 1-35, Nov. 2001.

[14] Jeremy C. Goh, Fuwei Jiang, Jun Tu, Yuchen Wang, “Can US economic variables predict the Chinese stock market?,” Pacific-Basin Finance Journal, Vol. 22, pp. 69-87, April. 2013. DOI: 10.1016/j.pacfin.2012.10.002

[15] Athanasios Koulakiotis, Apostolis Kiohos & Vassilios Babalos, “Exploring the interaction between stock price index and exchange rates: an asymmetric threshold approach,” Applied Economics, Vol. 47, No. 13, pp. 1273-1285, Jan. 2015. DOI: 10.1080/00036846.2014.990618

[16] MY, NQ. & SAYİM, M. “The Impact of Economic Factors on the Foreign Exchange Rates between USA and Four Big Emerging Countries: China, India, Brazil and Mexico,” International Finance and Banking, Vol. 3, No. 1, pp. 11-43, Feb. 2016. DOI: 10.5296/ifb.v3i1.9108

[17] Tsagkanos, Athanasios G. and Costas Siriopoulos. “A long-run relationship between stock price index and exchange rate: A structural nonparametric cointegrating regression approach,” Journal of International Financial Markets, Institutions and Money, Vol. 25, pp. 106-118, July. 2013. DOI: 10.1016/j.intfin.2013.01.008

[18] Schumaker, Robert P. and Hsinchun Chen. “Textual analysis of stock market prediction using breaking financial news: The AZFin text system,” ACM Trans. Information System, Vol. 27, No. 12 pp. 11-19, March 2009. DOI: 10.1145/1462198.1462204

[19] Gilbert, Eric and Karrie Karahalios. “Widespread Worry and the Stock Market,” International AAAI Conference on Weblogs and Social Media, Jan. 2010.

[20] Bollen, Johan, Huina Mao and Xiao-Jun Zeng. “Twitter mood

predicts the stock market," Journal of Computational Science. Vol. 2, No. 1, pp. 1-8, March. 2011. DOI: 10.1016/j.jocs.2010.12.007

[21] Lavrenko, Victor, Matthew D. Schmill, Dawn J. Lawrie, Paul Ogilvie, David D. Jensen and James Allan. "Mining of Concurrent Text and Time Series," KDD-2000 Workshop on Text Mining, Vol. 6, pp. 37-44, 2000.

[22] Hutto, Clayton J. and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," International AAAI Conference on Weblogs and Social Media, Vol. 8, pp. 216–225, Jan. 2015.

[23] Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," ArXiv: abs/1810.04805v2, May. 2019.

[24] Nohyoon Seong, Kihwan Nam, "Combining Macro-economical Effects with Sentiment Analysis for Stock Index Prediction," Entrue Journal of Information Technology, Vol. 16, No. 2, pp. 41-54, Aug. 2017.

[25] Ehsan Hoseinzade, Saman Haratizadeh, "CNNpred: CNN-based stock market prediction using a diverse set of variables," Expert Systems with Applications, Vol. 129, pp. 273-285, Sep. 2019. DOI: 10.1016/j.eswa.2019.03.029

[26] Akita, Ryo, Akira Yoshihara, Takashi Matsubara and Kuniaki Uehara. "Deep learning for stock prediction using numerical and textual information," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science, Vol. 15, pp. 1-6, June. 2016. DOI: 10.1109/ICIS.2016.7550882.

[27] Kalchbrenner, Nal, Edward Grefenstette and Phil Blunsom. "A Convolutional Neural Network for Modeling Sentences," ArXiv:1404.2188, April. 2014.

[28] Siami-Namini, Sima and Akbar Siami Namin. "Forecasting Economics and Financial Time Series: ARIMA vs. LSTM," ArXiv: abs/1803.06386, Mar. 2018.

[29] Mcnally, Sean, Jason T. Roche and Simon Caton. "Predicting the Price of Bitcoin Using Machine Learning," Euromicro International Conference on Parallel, Distributed and Network-based Processing, Vol. 26, pp. 339-343, March. 2018. DOI: 10.1109/PDP2018.2018.00060

[30] Lipton, Zachary Chase. "A Critical Review of Recurrent Neural Networks for Sequence Learning," ArXiv: abs/1506.00019, May. 2015.

[31] Christopher. Olah, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Authors

Euna Jang received the B.S. degree in Economics from University of California, San Diego, U.S. in 2015 and M.S. in Industrial Engineering from Korea University, Korea in 2020 respectively.

Euna Jang is currently a Student in the Department of Computer Science, Korea University. She is interested in AI, NLP, Big data analysis, Machine learning, Deep learning.

HoeRyeon Choi received the B.S. degree in Industrial Engineering from Dankook University, Korea in 1993. She finished the doctor coursework in Industrial Engineering at Korea University, Korea in 2005.

respectively. Dr. Choi currently working as a visiting professor in School of Industrial and Management Engineering at Korea university. She is interested in Big Data analysis, Machine Learning and AI Algrithms(Image Generation, Autonomous Driving)

HongChul Lee received the B.S. degree in Industrial Engineering from Korea University, Korea, in 1983, M.S. degree in Industrial Engineering from Texas Arlington University, U.S. in 1988 and he received Ph.D. degree

in Industrial Engineering from Texas A&M University, U.S. respectively. Dr. Lee joined the faculty of the Department of Industrial Management Engineering at Korea University, Seoul, Korea, in 1996. He is currently a Professor in the Department of Industrial Management Engineering, Korea University. He is interested in Artificial Intelligence, Manufacturing Engineering System, and Simulation.