노이즈 데이터에 따른 애플 주가 예측 정확도 분석

(LSTM & Sentence Sentiment Classification)

**선풍기 바람 쐬는 어피치**

# Introduction

해당 프로젝트는 2020년 1월 1일 ~ 2023년 12월 31일까지의 애플(APPL) 주가 데이터와 노이즈 데이터 그리고 LSTM 모델을 통해 2024년 1월 1일 ~ 2024년 1월 31일의 종가를 예측하는 프로젝트입니다.

해당 프로젝트를 통해 노이즈 데이터를 모델에 학습시킨다면 주가 데이터만을 학습했을 경우보다 얼마나 더 정확한 예측이 가능한지에 대해 알 수 있을 것으로 기대합니다.  
더 나아가 노이즈 데이터로 인한 폭락을 대비할 방법을 수립하는 것에 대해서도 기대하고 있습니다.

해당 프로젝트에서 사용할 LSTM 모델의 경우 기존 RNN 대비 장기 의존성 문제를 해결하여 긴 기간을 학습하고 예측해야 하는 주가 예측에 적합한 모델입니다.

그럼에도 불구하고 주가를 결정 짓는 요소는 과거 주가 데이터 뿐만이 아닌 노이즈 데이터(e.g. 경제 지표, 정치적 사건, 기업의 성과 발표 및 불확실한 미래의 갑작스러운 사건 등) 또한 무시하지 못할 요소입니다.   
때로는 이 노이즈 데이터가 주가에 큰 영향을 미치기도 합니다.

그렇기에 해당 프로젝트에서는 주가 데이터만을 사용한 LSTM 모델과 주가 데이터와 노이즈 데이터를 까지 사용한 LSTM 모델 간의 정확도 분석을 통해  
노이즈 데이터를 학습시킨다면 주가를 더 정확하게 예측하고 또한 예상치 못한 폭락에 대해 대비가 가능하도록 연구해볼 목적입니다.

많은 주식 중 애플을 선정한 이유와 애플 주식에 대한 간단한 소개는 아래와 같습니다.  
애플(APPL)은 미국 주식 시장의 대표적인 우량주 중 하나이며 24년 3월 18일(UTC) 기준 173.72$를 기록하고 있습니다.

또한 지난 1년간 최고 199.62$( ▲14.9%)  
최저 154.15$(▼12.3%)를 기록하는 등 비교적 안정적인 예측이 가능할 것으로 기대하고 있습니다.  
특히 해당 프로젝트에서 데이터로 사용할 기간 에는 적당한 수준의 상승과 하락이 반복되며   
전일 대비 최고 12.0% 상승 그리고 -12.9% 하락하는 날도 존재하므로 해당 일에 대해 노이즈 데이터는 어떤 양상을 띠고 있었는지에 대해 연구하기에 적합한 주식입니다.

# Problem Formulation

해당 프로젝트에서는 주가 데이터를 활용하여 LSTM 모델을 통해 주가를 예측합니다.

LSTM을 통한 주가 예측의 한계는 주가는 시계열 데이터이지만 과거 주가 외의 노이즈 데이터에 의한 영향이 크다는 점입니다. 이를테면 특정 기업의 CEO가 SNS에 기행을 펼치면 사람들은 CEO의 이상한 행동을 보며 해당 기업의 신뢰도를 잃게 되고 결과적으로 이는 주가에 악영향을 미치는 결과가 될 것입니다.  
위의 예시처럼 미래의 주가는 무조건 과거의 주가에만 영향을 받지 않으며 학습 데이터 이외의 노이즈 데이터에 주가 영향을 받는다는 점이 LSTM 모델을 통한 주가 예측의 한계입니다.

그렇기에 해당 프로젝트에서는 주가 데이터 이외의 노이즈 데이터 또한 함께 학습합니다.  
여기서 노이즈 데이터는 전체적인 시장 흐름을 파악하기 위한 Index 지수와 애플 기업 관련된 뉴스 기사입니다.  
여기서 뉴스 기사의 경우 제목만 활용하며 Sentiment Classification을 통해 기업에 대해 기사 내용이 적대적인지 중립적인지 우호적인지에 대해 정수로 표현하여 함께 학습합니다.  
이로써 기존의 LSTM 주가 예측의 부족한 데이터를 보충해 줌으로써 조금 더 정확한 예측이 가능해질 것으로 기대할 수 있습니다.

# Architecture

## Dataset & Featrue:

주가 데이터로는 애플 주식의 2019년~2024년 데이터를 수집하여 사용할 계획입니다.

해당 데이터에는 시가, 저가, 고가, 종가, 수정 종가, 거래량 정보가 포함되어 있으며

2019년 데이터를 활용하여 2020년 데이터에 20일, 60일, 120일 이동평균선을 추가하여 train-data로 활용할 예정입니다.

추가로 거래량의 단위가 높기 때문에 정규화 작업을 필수로 진행해야 합니다.

주가 데이터 수집 방법으로는 주가 정보 제공 사이트로 잘 알려진 야후 파이낸스의 Historical data를 다운받아서 준비할 예정입니다.

Noise data로는 뉴스 기사 제목, index 지수인 NASDAQ, S&P 500 그리고 Dollar Index를 사용할 계획입니다.

NASDAQ 지수는 애플을 포함한 주요 기술 기업들이 상장되어 있으며 해당 지수는 시장 전체의 심리를 반영한다고 볼 수 있습니다.

S&P 500 지수는 미국 시장의 500대 대표 기업을 포함하고 있으며 이 또한 애플 INC를 포함한 시장 전체 심리를 반영한다고 볼 수 있습니다.

US Dollar Index는 6개국(유럽, 일본, 캐나다, 스웨덴, 스위스)의 통화와 비교하여 미국의 달러 평균 가치를 나타내는 지수입니다.

해당 지수에 따라 투자자의 심리를 변화할 수 있으며 이는 미국 주식 시장에도 영향을 끼친다고 판단할 수 있습니다.

위 세가지 index 지수의 수집 방법에 대해서는 애플 주식 데이터와 마찬가지로 야후 파이낸스에서 수집할 예정입니다.

마지막으로 뉴스 기사 제목의 경우 제목은 본문의 내용을 함축하는 의미이므로 제목을 감정 분석하여 애플에 우호적인지 중립적인지 적대적인지 점수를 산출하여 LSTM 모델 학습에 추가할 예정입니다.

수집 방법으로는 웹크롤링을 활용하며 구글을 통해 "apple stock" 키워드로 검색하여 출력되는 뉴스 기사들의 제목을 수집할 계획입니다.

## Model:

## 해당 프로젝트의 학습 데이터인 주가 데이터는 시간에 따른 관측 데이터이며 각 원소뿐만이 아닌 시간적인 순서 또한 중요한 데이터이기 때문에 순환 신경망을 선택할 것이며

## 이 중에서도 장기 의존성 문제를 해결할 수 있는 LSTM(Long Short Term Memory) 모델을 선택할 예정입니다.

## 주가 데이터의 경우 LSTM을 활용하며 Index 지수 또한 주가 데이터와 동일한 종류의 데이터이기 때문에 똑같이 LSTM을 활용할 예정이며 뉴스 제목에 대해서 Sentence Sentiment Classification을 진행할 모델 또한 LSTM을 활용할 것으로 예상합니다.

## 하지만 Sentence Sentiment Classification에 대해서는 LSTM 외에 RNN, BERT, Transformer 모델 또한 고려하고 있습니다.

## 최종적으로 모델의 입력층의 경우 애플 주가 데이터, Index 지수 데이터, 뉴스 제목 데이터 각각 세 가지의 입력을 받아서 학습을 진행한 뒤 출력층은 애플 주식의 종가를 출력합니다.

## Loss Function:

## 해당 프로젝트는 애플 주식의 종가를 예측하는 회귀 문제 이므로 예측한 종가와 실제 종가의 차이를 측정할 수 있는 Mean Squared Error(MSE)를 사용할 것입니다.

## 또한 훈련 중 과적합이 발생하는 것을 방지하기 위해 Hidden Layer에 Drop-out층을 추가할 것이며 그 외에도 MSE에 적용할 수 있는 규제인 L2 규제 또한 고려해 볼 수 있을 것으로 예상됩니다.

## Train data set의 일부를 validation data으로 사용하여 하이퍼 파라미터 튜닝을 진행할 예정이며 Optimizer는 범용적으로 좋은 성능을 보이는 Adam Optimizer를 선택할 것입니다.

## Evaluation:

해당 프로젝트는 2가지의 평가지표를 활용합니다. 첫번째로 Mean Absolute Error(MAE) 입니다. 예측과 실제의 차이의 절대값을 통해 해당 모델의 오차가 어느 정도 수준인지 직관적으로 가늠할 수 있습니다.

두번째로 Mean Absolute Percentage Error(MAPE) 를 활용합니다. MAPE 지표를 활용함으로써 모델의 예측값의 절대 오차의 평균을 백분율로 나타내어 조금 더 직관적인 평가를 확인할 수 있습니다. MAPE의 특성상 실제 값이 0에 가까울 경우 MAPE 수치가 발산할 수도 있으나 실제 주가의 가격이 0에 수렴하지 않기 때문에 사용에는 적합할 것으로 판단됩니다.