Proposal Report

2020170837 최원준 | 2021320328 장지웅 | 2022320157 노규주

1 Problem Definition

본 프로젝트는 영향력 있는 인물들의 SNS 발언 데이터를 수집·분석하여, **'발언과 국제 경제 지표 변동 간의 연관성을 규명하고 이를 예측할 수 있는 머신러닝 기반 모델을 구축'**하는 것을 목표로 한다. 특히 인물별·발언별 영향력을 정량화하여 분류하거나 클러스터링함으로써, **발언이 가지는 시장 파급 효과를 체계적으로 모델링**하고자 한다.

2 Interest and Significance of the Problem

정치인과 기업인의 발언은 국제 경제 지표와 금융시장에 즉각적이고 큰 영향을 미친다. 트럼프 전 대통령의 트윗, 일론 머스크의 SNS 발언, 제롬 파월 연준 의장의 연설 등은 금리, 환율, 주가 변동을 촉발한 대표적 사례다. 특히 SNS 확산으로 발언이 시장에 전달되는 속도는 과거보다 훨씬 빨라졌으며, 트위터(X)와 같은 플랫폼은 글로벌 금융시장과 즉시 연결된다. 이처럼 발언 직후 자산 가격이 곧바로 반응하는 경우가 많아, 기존 거시경제 지표만으로는 설명하기 어려운 단기 변동성과 새로운 정보 흐름을 형성한다.

금융시장은 본래 급등락과 heavy tails 등으로 예측이 어렵지만, 발언은 단기적 시장 움직임을 가늠할 수 있는 중요한 신호가 된다. 따라서 발언과 시장 지표 간 관계를 데이터 기반으로 분석하고 이를 머신러닝 모델에 적용하면 기존 모델이 놓칠 수 있는 단기 패턴과 새로운 흐름을 포착할 수 있다. 더 나아가 발언 기반 리스크를 정량화하면 환율·금리·원유 가격 등 주요 지표의 예측과 위험 관리, 정책 효과 분석, 시장 안정 장치 설계에 활용할 수 있다. 또한 인물별 발언 특성을 분류·클러스터링하여 영향력을 체계적으로 측정하면 발언과 경제 지표 간 관계를 보다 명확히 이해할 수 있다.

3 Machine Learning Methods and Rationale

본 프로젝트는 특정 인물의 SNS 발언이라는 **unstructured data**(e.g., text)를 입력으로 받아 연속적 금융 지표(e.g., S&P500, 환율)의 변동을 예측하는 것을 목표로 한다. 이는 ***Supervised Learning task*** 중 하나인 ***Regression task*** 로 설정되며, 예측 정확도를 높이고 결과를 explainable하게 제시하기 위해 다양한 보조적 기법들을 활용한다. 이를 위해 다음과 같은 단계별 머신러닝 방법론을 적용하고자 한다.

3.1 Data preprocessing and Feature Engineering

SNS 발언 텍스트 데이터를 입력 받아 ***meaningful vector representation***으로 변환한다. 이를 위해 텍스트를 토큰화하고 **TF-IDF, word embedding, pre-trained language model**(e.g., Word2Vec, BERT)을 활용해 벡터화한다. 추가적으로 **LLM**을 이용한 주제 태깅과 Sentiment Analysis를 통한 긍정·부정 점수를 포함하여 발언의 의미와 맥락을 반영한다.

3.2 Predictive Modeling

전처리된 특징 벡터를 입력으로 일정 기간 이후 금융 지표 변동을 예측하는 Machine Learning 기반 모델을 구축한다. 모델 학습 과정에서는 validation set과 test set을 통해 성능을 검증하고, generalization 성능을 확보한다. 기존의 전통적 금융 이론이나 단순 통계 모델만으로는 SNS 발언과 금융 지표 변동 간의 복잡한 상호작용을 설명하기 어렵다. 이에 따라 본 프로젝트는 비정형 데이터 처리, 비선형 관계 학습, 시계열 데이터 분석이 가능한 머신러닝 기법을 적용한다. 또한 단순한 선형 관계에서 시작해 점차 비선형 및 시계열 패턴으로 확장하는 단계적 접근을 통해, 예측 성능과 해석 가능성을 동시에 확보하고자 한다.

- **Classical Statistical and ML Methods**
  Multiple Linear Regression을 적용해 변수 간 선형 관계를 파악하고 해석 가능성을 확보한다. 이어 **Ridge/Lasso Regression, ARIMA, SVR** 등 전통적 기법을 활용해 baseline을 넘어 데이터와 비선형 관계를 다룬다. 이러한 모델들은 단순하면서도 해석이 용이해 이후 발전된 모델과 비교할 기준점을 제공한다.

- **Ensemble Methods**
  Random Forest, Gradient Boosting(e.g., XGBoost, LightGBM, CatBoost)은 비선형성과 변수 간 상호작용을 효과적으로 학습한다. 특히 tabular 형태의 금융 데이터에서 강력한 성능을 보이며, feature importance와 같은 해석 지표를 제공할 수 있다는 점에서 유용하다.
- **Neural Network Models**
  MLP는 다양한 피처의 비선형 관계를 학습하는 데 적합하다. **LSTM**과 **GRU**는 발언과 금융 지표 간 시계열 흐름과 단기·장기 의존성을 효과적으로 포착하며, **Transformer**는 self-attention을 통해 복잡한 상호작용과 장기 패턴을 학습한다. 이러한 딥러닝 모델들은 전통적 기법이 놓치는 비선형성, 지연 효과, 변수 간 상호작용을 반영해 금융 지표 변동을 정밀하게 예측할 수 있다.
- **Classification & Clustering**
  본 프로젝트의 핵심은 Regression이지만, 결과의 explainability를 높이기 위해 Classification과 Clustering 기법도 보조적으로 활용한다. **Classification**(e.g., Random Forest Classifier, SVM)으로 금융 지표의 방향성(상승/하락)을 단순화된 위험 신호로 제시하고, **Clustering**(e.g., K-means, DBSCAN)으로 유사한 발언을 그룹화하여 어떤 유형의 발언이 금융시장에 큰 영향을 주는지를 패턴 차원에서 분석한다.
- **Advanced Deep Learning and Domain-Specific Models**
  본 프로젝트의 범위를 넘어서는 심화 연구 단계에서는 **시계열 Transformer 모델**들과 **FinBERT, N-BEATS, Temporal Fusion Transformer**와 같은 금융 특화 딥러닝 모델을 적용하여 예측 성능을 한층 고도화할 수 있다.

4 Project timeline & Team structure

- 문제 정의 및 데이터 수집: 09.24 ~ 10.08
- ML 모델 설계 및 구현 : 10.23 ~ 10.30
- 중간 보고서 작성 및 제출 : 10.30 ~ 11.04
- 모델 평가 및 개선 : 11.04 ~ 12.03
- 최종 보고서 작성 : 12.03 ~ 12.16

5 Team structure

- **노규주**: 데이터 수집 및 전처리, 모델 구현, 성능 평가 및 분석, 결과 해석 및 시각화
- **장지웅**: 데이터 수집 및 전처리, 모델 구현, 성능 평가 및 분석, 결과 해석 및 시각화
- **최원준**: 데이터 수집 및 전처리, 모델 구현, 성능 평가 및 분석, 결과 해석 및 시각화

6 Expected results

본 프로젝트의 최종 목표는 다음과 같은 세 가지 핵심 결과물을 구현하는 것이다. 각 결과물은 제안된 문제에 대한 구체적인 해결 방안을 제시하고, 데이터 기반 모델링 역량을 증명하는 데 초점을 둔다.

1. **SNS 발언 기반 변동성 예측 모델 구현**
   주요 인물의 SNS 발언, 이미지 등을 입력받아 S&P 500, 환율과 같은 경제 지표의 단기 변동성을 예측하는 ML 모델을 구현한다. 모델의 성능을 객관적으로 측정(e.g., accuracy, f1-score) 하고 그 결과를 분석한다.

2. **발언 영향력 정량화 및 분류 시스템 구축**
   NLP 기술을 적용하여 개별 발언의 잠재적 시장 파급력을 보여주는 'Impact Score'를 산출하는 시스템을 구축한다. 나아가, 유사한 주제의 발언들을 자동으로 묶어주는 클러스터링 기능을 구현하여, 특정 주제(e.g., 금리, 무역)가 갖는 리스크를 분류하고 분석할 수 있도록 한다.

3. **데이터 시각화를 통한 핵심 패턴 분석**
   분석된 데이터를 시각화하여 특정 인물과 발언이 어떤 시장에 유의미한 영향을 미치는지 그 패턴을 명확히 보여준다. '파월의 발언은 금리에, 머스크의 트윗은 기술주에 영향을 준다'와 같은 핵심 관계를 차트나 그래프로 도출하여, 모델의 분석 결과를 직관적으로 이해할 수 있도록 돕는다.