

SETION 04

AI 부트캠프 10
박민경

—

Financial Sentiment Analysis

목차

CONTENTS

- 1 주제 및 가설 설정
- 2 데이터 전처리
- 3 모델 학습 및 평가
- 4 결론 및 회고

주제 및 가설 설정

데이터 & 주제 선정

- 📎 금융 시장은 투자자의 심리와 행동에 의해 민감하게 반응하는 시장
- 📎 finance와 연관된 문장에서 어떤 감정이나 의견을 가지고 있는지를 분석할 것
- ▶ 후에 이것이 어떻게 시장에 작용하는지에 대한 프로젝트로 발전시켜 경제 흐름이나 시장 전반에 대한 이해를 도울 수 있도록 하고자 함

주제 및 가설 설정

가설 설정

가설

▶ LSTM 모델보다 BERT 모델의 성능이 감성 분석(sentiment analysis)에 더 뛰어날 것이다.

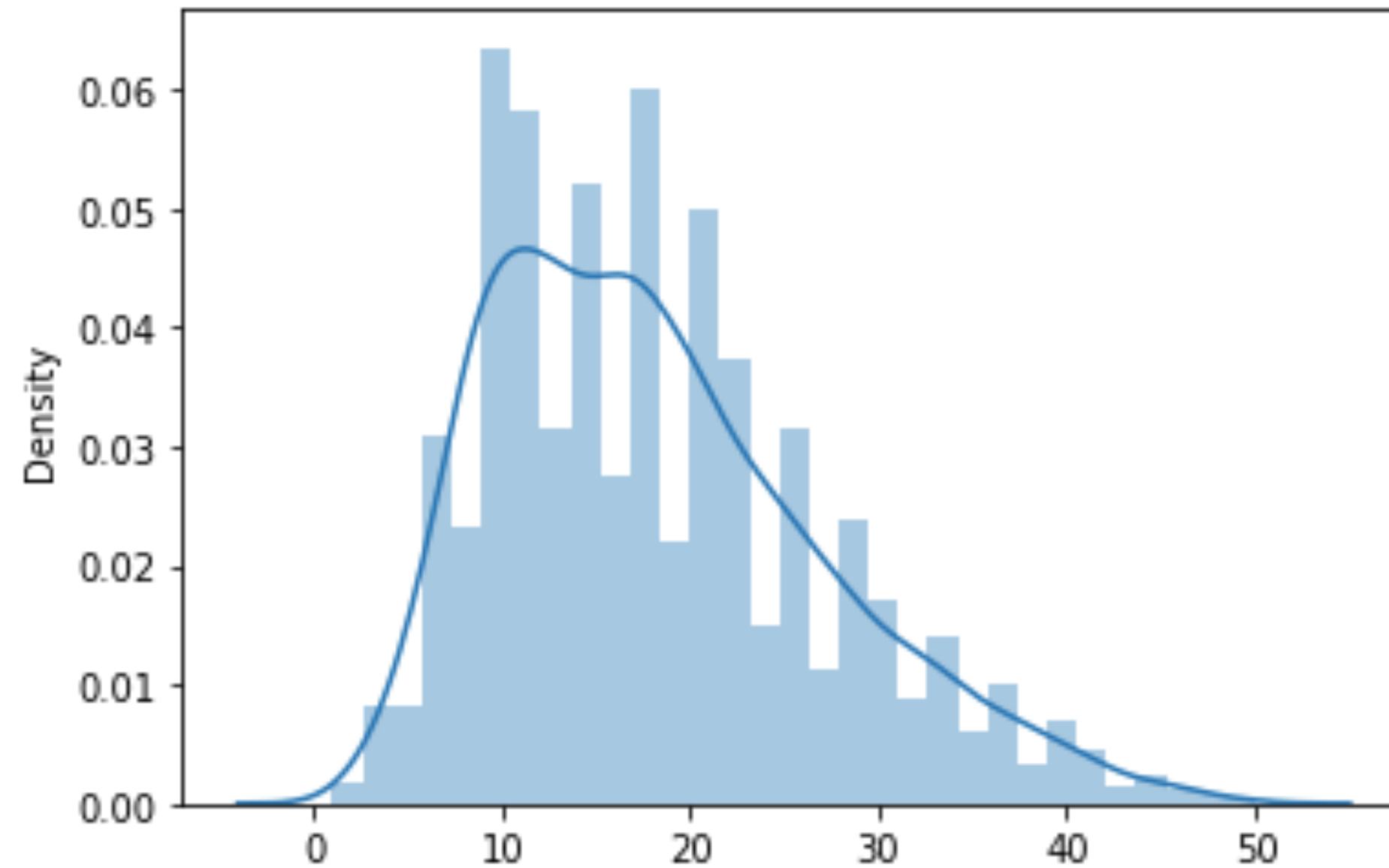
= 본 프로젝트의 문제(financial sentiment analysis)에는 BERT 모델이 더 적합하다.

02 데이터 전처리

- 📎 http나 @ 등의 특수 기호 제거해 텍스트를 정제
- 📎 토큰 수를 줄이기 위해 정규 표현식을 사용하여 대/소문자와 숫자를 제외한 문자들을 제거
- 📎 토큰 길이 중 최장 길이에 맞춰 padding 진행
- 📎 Sentiment: neutral, negative, positive로 이루어짐
 - ▶ LabelEncoder로 인코딩

02 데이터 전처리

토큰 길이 분포도 📉



Model: "sequential_2"

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 52, 64)	797440
conv1d_4 (Conv1D)	(None, 48, 64)	20544
conv1d_5 (Conv1D)	(None, 44, 32)	10272
max_pooling1d_2 (MaxPooling 1D)	(None, 22, 32)	0
lstm_2 (LSTM)	(None, 32)	8320
dropout_2 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 3)	99

Total params: 836,675
Trainable params: 836,675
Non-trainable params: 0

- 📎 과적합 방지를 위해 dropout과 학습률 감소를 적용하고, kernel regularizer와 bias regularizer 설정
- 📎 optimizer: Adam
- 📎 평가지표: loss, 정확도, f1 score
- 📎 epoch=20, batch_size=32 설정

📎 20번째 epoch에서 loss: 1.0820, accuracy: 0.5344, f1 score: 2.1946

📎 모델 평가한 결과: loss: 1.1033, accuracy: 0.5708, f1 score: 2.0136

▶ 모델 평가 시 도출된 지표들 중 **loss, f1 score**는 학습 결과 보다 근소하게 감소했고, **정확도**는 근소하게 증가함

- 📎 해당 데이터를 BERT로 학습할 때, tensorflow로 진행할 수 없어 Pytorch로 진행함
- 📎 fine-tuning 진행해 모델 구조와 가중치를 조정
- 📎 epoch=20, batch_size=32, optimizer=AdamW로 설정
- 📎 과적합 방지를 위해 학습률 감소 설정

📎 학습 결과: epoch=20에서 loss=0.10, validation accuracy: 0.76

📎 test dataset으로 평가한 결과: accuracy: 0.7545


▶ 학습 시와 평가 시의 정확도 간의 차이가 거의 없음

결론

📎 정확도 측면에서 LSTM보다 BERT가 더 높음
= financial sentiment analysis에 더 적합하다는 것을 보여줌

한계 및 발전 방향

- 📎 BERT에 대해 이해하지 못 한 부분이 많고, 코드가 어려워서 많은 것을 시도해보지 못 함
- 📎 BERT에 대해 좀 더 공부를 한 후, CV와 하이퍼 파라미터 시도 + FinBERT로 다시 시도해볼 것



감사합니다 :D