2023-2 경제학연습 2차 발표

Korean Financial Sentiment Embedding

전지훈

https://github.com/JihoonBrianJun/KoFinEmb

Dataset

title	label
♥나지금 호수공원서 주식점 봐주고있음 ♥	[하락]
이거 저도 갖고 있지만 안 좋은 추세네요	[하락]
셀트리온에 큰절 10번 올려라 재용아.	[상승]
"경기침체땐 현금이 최고"...低PCR주 뜬...	[하락]
기레기들이 잔뜩 분위기 조성 실컷 해놓고...	[하락]
삼성물산 도적단 IR 직통번호 02-34...	[하락]
근데 상폐론자들. 상폐 왜되는거야???	[하락]
항체표적암기술로삼성손잡을수밖에없는이유공개	[하락]
곧 17만원온다. 손절하고 기다려라.	[하락]
원전30조?ㅋ여기 영업이익마이너스30조	[상승]
영화한편 보고 탈원전한다는 저능통뽑아놓고...	[상승]
LG화학 시총도 같은 기간 57%나 늘어...	[상승]
오늘을 기점으로 18만이 바닥이 되었으면...	[상승]
▪분식삼바 그네 친일빠들의 공감쇼 ㅋ	[상승]
금융위원장 은성수를 해임하여 주십시요	[하락]
새벽에 금수들이 움직인다더니!! z...	[하락]
앞으로의 방향성은 30% 폭등이...!!	[상승]
18만원에서20만원 사이 개스코 한달내내...	[상승]
내가 이개잡주를 왜담앗을까.. 하..	[하락]
이젠 하다하다 개 한테까지 세금먹이네 ㅋ...	[하락]
개같은주식이네 코스피상승할동안 한번도오른...	[상승]
한전 특징: 투자는 엄청 말하면서 정작 ...	[하락]
호가창에 13만에 500억원있네 지린다	[상승]
게10판알바새기야 김종갑이 누구냐? 니애...	[하락]
文 잘한다 36.6% vs 못한다 60....	[하락]
히융 익절예상가 30인데 올해안에 올까?	[하락]
내릴 때 사고 오를 때 파는게 기본인데	[하락]
개잡주!!!!!!!!!!!!!!!! ㅅ ...	[하락]

- 네이버증권 종목토론방 데이터
  - 개인투자자 Sentiment 텍스트 데이터 :
    - KOSPI 시가총액 상위 20개 종목 (2022.12 기준)
      - 2017년 이후 상장 종목 제외하고 상위 20개 계산
    - 각 종목별 5.5년치 데이터 (2017.6 ~ 2022.12)
    - 게시글 제목만 활용
      - 본문 및 댓글 포함은 추후 구현
  - Target Label 데이터 :
    - 해당 게시글 작성일 종가의 전일 대비 [상승] / [하락]

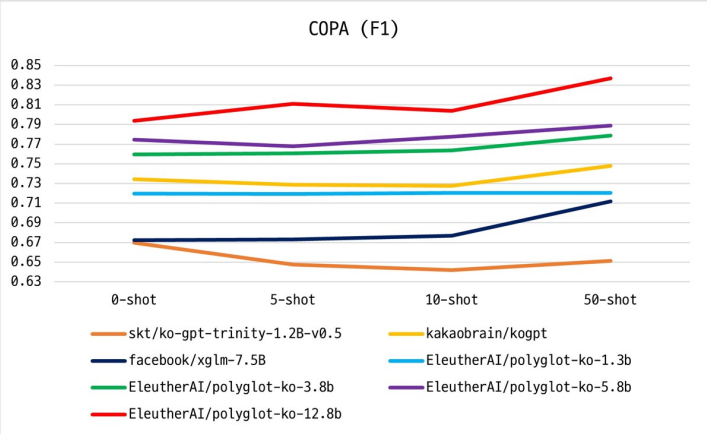
Dataset

title	label
♥나지금 호수공원서 주식점 봐주고있음 ♥	[하락]
이거 저도 갖고 있지만 안 좋은 추세네요	[하락]
셀트리온에 큰절 10번 올려라 재용아.	[상승]
"경기침체땐 현금이 최고"...低PCR주 뜬...	[하락]
기레기들이 잔뜩 분위기 조성 실컷 해놓고...	[하락]
삼성물산 도적단 IR 직통번호 02-34...	[하락]
근데 상폐론자들. 상폐 왜되는거야???	[하락]
항체표적암기술로삼성손잡을수밖에없는이유공개	[하락]
곧 17만원온다. 손절하고 기다려라.	[하락]
원전30조?ㅋ여기 영업이익마이너스30조	[상승]
영화한편 보고 탈원전한다는 저능통뽑아놓고...	[상승]
LG화학 시총도 같은 기간 57%나 늘어...	[상승]
오늘을 기점으로 18만이 바닥이 되었으면...	[상승]
▪분식삼바 그네 친일빠들의 공감쇼 ㅋ	[상승]
금융위원장 은성수를 해임하여 주십시요	[하락]
새벽에 금수들이 움직인다더니!! z...	[하락]
앞으로의 방향성은 30% 폭등이...!!	[상승]
18만원에서20만원 사이 개스코 한달내내...	[상승]
내가 이개잡주를 왜담앗을까.. 하..	[하락]
이젠 하다하다 개 한테까지 세금먹이네 ㅋ...	[하락]
개같은주식이네 코스피상승할동안 한번도오른...	[상승]
한전 특징: 투자는 엄청 말하면서 정작 ...	[하락]
호가창에 13만에 500억원있네 지린다	[상승]
게10판알바새기야 김종갑이 누구냐? 니애...	[하락]
文 잘한다 36.6% vs 못한다 60....	[하락]
히융 익절예상가 30인데 올해안에 올까?	[하락]
내릴 때 사고 오를 때 파는게 기본인데	[하락]
개잡주!!!!!!!!!!!!!!!! ㅅ ...	[하락]

- 네이버증권 종목토론방 데이터
  - 개인투자자 Sentiment 텍스트 데이터 :
    - 전체 데이터 중 3만건 활용
      - Label이 [상승]인 데이터 1.5만건 (Random sampling)
      - Label이 [하락]인 데이터 1.5만건 (Random sampling)
    - Train 데이터 (99%)와 Test 데이터 (1%)로 분할

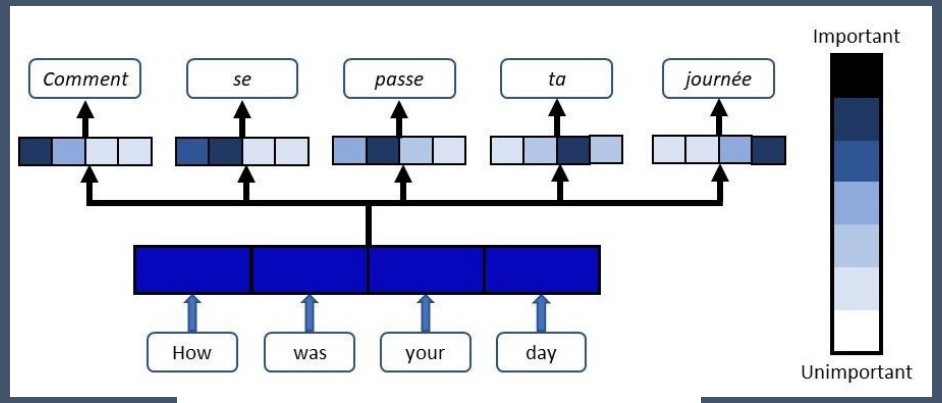
Model

Model	params	0-shot	5-shot	10-shot	50-shot
skt/ko-gpt-trinity-1.2B-v0.5	1.2B	0.6696	0.6477	0.6419	0.6514
kakaobrain/kogpt	6.0B	0.7345	0.7287	0.7277	0.7479
facebook/xglm-7.5B	7.5B	0.6723	0.6731	0.6769	0.7119
EleutherAI/polyglot-ko-1.3b (this)	**1.3B**	**0.7196**	**0.7193**	**0.7204**	**0.7206**
EleutherAI/polyglot-ko-3.8b	3.8B	0.7595	0.7608	0.7638	0.7788
EleutherAI/polyglot-ko-5.8b	5.8B	0.7745	0.7676	0.7775	0.7887
EleutherAI/polyglot-ko-12.8b	12.8B	0.7937	0.8108	0.8037	0.8369



- 한국어 언어모델 : polyglot-ko 1.3b
  - 최신 고성능 언어모델들은 대부분 영어 기반
    - ChatGPT, GPT-4, Llama2 등
    - 학습 데이터에 한국어도 포함되어 있기는 하나, 극히 일부
      - 모델에 따라 0.01~0.06% 정도
  - 최신 한국어 언어모델들은 오픈소스가 아닌 것들이 많음
    - 네이버 하이퍼클로바X 등
  - 현재 오픈소스 한국어 모델 중 가장 성능 좋은 것으로 알려짐

Model : How It Works



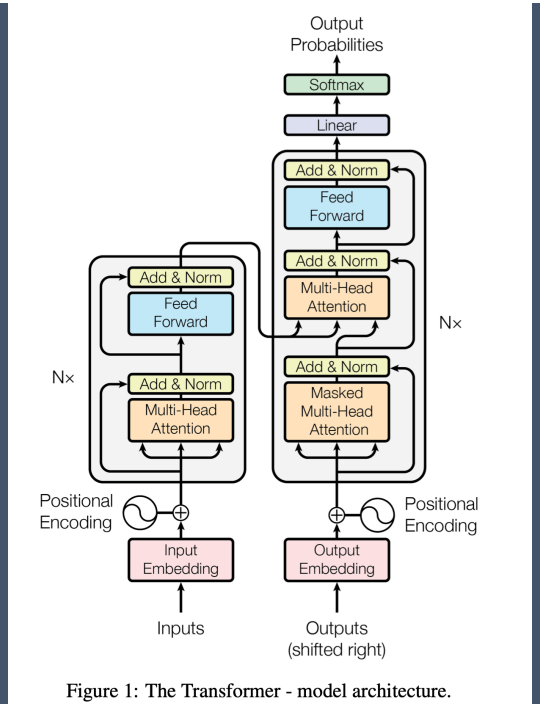


Figure 1: The Transformer - model architecture.

- Transformer 레이어를 반복적으로 통과
  - 우선 주어진 텍스트를 토큰화하여 모델에 투입
    - Input : $Batch\_Size \times Sequence\_length$ 차원
  - 모델은 각 토큰들을 4096차원 벡터로 형상화
    - Output : $Batch\_Size \times Sequence\_length \times 4096$ 차원
  - 4096차원 벡터 결정(계산) 방법
    - Sequence 내 연관도 높은 토큰 간 내적 값 높게!
  - 토큰의 가짓수 (Token ID)
    - 모델마다 상이하나, 이 모델은 30003개

Model : How It Works

```python
for step, batch in enumerate(tqdm(train_dataloader,colour="blue", desc=f"Training Epoch{epoch}")):
    for key in batch.keys():
        if config.enable_fsdp:
            input_ids = batch["input_ids"].to(local_rank)
            labels = batch["labels"].to(local_rank)
            attention_mask = batch["attention_mask"].to(local_rank)
        else:
            input_ids = batch["input_ids"].to('cuda:0')
            labels = batch["labels"].to('cuda:0')
            attention_mask = batch["attention_mask"].to('cuda:0')

    outputs = model(input_ids=input_ids, attention_mask=attention_mask)
    lm_logits = outputs[0]    # (BS, Max_Seq_Len, vocab_size)

    loss = None

    logits = lm_logits[:,:-1,:].contiguous().view(-1, lm_logits.size()[-1])    # (BS * (Max_Seq_Len-1), vocab_size)
    labels = labels[:,1:].contiguous().view(-1).to(logits.device)    # (BS * (Max_Seq_Len-1))
    loss_function = CrossEntropyLoss()
    loss = loss_function(logits, labels)
```

- Transformer 결과물(토큰별 4096차원)을 30003차원(전체 토큰 종류 수)으로 Projection
  - Sequence 내 각 Token마다 다음 Token ID를 예측
    - 30003차원 벡터 내에서 k번째 element가 가장 크면, 다음 Token ID를 k로 예측
  - 각 batch별로 예측값 및 실제 정답들을 종합해 Loss를 계산 후, 모델 최적화 진행 (학습 시)
    - 정답을 맞추도록 30003차원 벡터 값들의 조정이 일어남 (➔ 4096차원 벡터로도 역전파)

Model Training

```
Example dataset..
Input: 주식 종목 토론방에 [제목]의 글이 올라온 다음 날, 주가는 어떻게 되었을까?
       [상승] 혹은 [하락] 중 하나로 답하여라.

       [제목]: 장마감 이후 올라온 공시 자료를 보니, 분기 실적이 예상치를 크게 밑도네요ㅠ
       [정답]: [하락]

       [제목]: 간밤에 미국 연준의 깜짝 금리 인하 결정에 힘입어 미국 증시 폭등 중이네요. 내일 한국 증시도 기대해 봅니다!
       [정답]: [상승]


[제목]: ♥나지금 호수공원서 주식점 봐주고있음 ♥
[정답]:
```

- 2-Shot Finetuning
  - 사전학습된 모델을 가져와서, 학습 데이터셋을 이용해 모델 미세조정
  - Input : 지시 사항 + 올바른 예시 2개 ([제목]&[정답]) + 종목토론방 게시글 제목
    - 예시 2개는 가상의 이상적인 형태로 직접 만들어서 넣어줌
  - 문맥 학습 (In-Context Learning)을 통해 모델이 올바른 정답 도출하도록 자연스레 유도

Similar Work

I. Try our model (FinGPT v3)

Code:

```
from transformers import AutoModel, AutoTokenizer, AutoModelForCausalLM, LlamaForCausalLM, LlamaTokenizerFast
from peft import PeftModel  # 0.5.0

# Load Models
base_model = "NousResearch/Llama-2-13b-hf"
peft_model = "FinGPT/fingpt-sentiment_llama2-13b_lora"
tokenizer = LlamaTokenizerFast.from_pretrained(base_model, trust_remote_code=True)
tokenizer.pad_token = tokenizer.eos_token
model = LlamaForCausalLM.from_pretrained(base_model, trust_remote_code=True, device_map = "cuda:0", load_in_8bit =
model = PeftModel.from_pretrained(model, peft_model)
model = model.eval()

# Make prompts
prompt = [
'''Instruction: What is the sentiment of this news? Please choose an answer from {negative/neutral/positive}
Input: FINANCING OF ASPOCOMP 'S GROWTH Aspocomp is aggressively pursuing its growth strategy by increasingly focu
Answer: ''',
'''Instruction: What is the sentiment of this news? Please choose an answer from {negative/neutral/positive}
Input: According to Gran , the company has no plans to move all production to Russia , although that is where the
Answer: ''',
'''Instruction: What is the sentiment of this news? Please choose an answer from {negative/neutral/positive}
Input: A tinyurl link takes users to a scamming site promising that users can earn thousands of dollars by becomi
Answer: ''',
]

# Generate results
tokens = tokenizer(prompt, return_tensors='pt', padding=True, max_length=512)
res = model.generate(**tokens, max_length=512)
res_sentences = [tokenizer.decode(i) for i in res]
out_text = [o.split("Answer: ")[1] for o in res_sentences]

# show results
for sentiment in out_text:
    print(sentiment)
```

- 최근 가장 널리 사용되는 모델 Finetuning 방법

Stanford Alpaca

Eval Result

```
주식 종목 토론방에 [제목]의 글이 올라온 다음 날, 주가는 어떻게 되었을까?
        [상승] 혹은 [하락] 중 하나로 답하여라.

        [제목]: 장마감 이후 올라온 공시 자료를 보니, 분기 실적이 예상치를 크게 밑도네요ㅠ
        [정답]: [하락]

        [제목]: 간밤에 미국 연준의 깜짝 금리 인하 결정에 힘입어 미국 증시 폭등 중이네요. 내일 한국 증시도 기대해 봅니다!
        [정답]: [상승]


[제목]: 안티들 밤새 발버둥..하락에 베팅했으니...
[정답]: [하락]<|endoftext|>
Gold Answer: [상승]
```

```
주식 종목 토론방에 [제목]의 글이 올라온 다음 날, 주가는 어떻게 되었을까?
        [상승] 혹은 [하락] 중 하나로 답하여라.

        [제목]: 장마감 이후 올라온 공시 자료를 보니, 분기 실적이 예상치를 크게 밑도네요ㅠ
        [정답]: [하락]

        [제목]: 간밤에 미국 연준의 깜짝 금리 인하 결정에 힘입어 미국 증시 폭등 중이네요. 내일 한국 증시도 기대해 봅니다!
        [정답]: [상승]


[제목]: 60만원이 다시 올 수 밖에 없는 이유
[정답]: [상승]<|endoftext|>
Gold Answer: [하락]
```

- Test 데이터 300건 중 148건 (49%) 정도만 정답
  - 정답이 [상승] / [하락] 둘 중 하나인 점 감안할 때, 거의 랜덤에 가까운 성능
  - 반어적 표현, 과신(Overconfidence) 등이 포함된 [제목]에 특히 취약한 경향

Eval Result

```
주식 종목 토론방에 [제목]의 글이 올라온 다음 날, 주가는 어떻게 되었을까?
        [상승] 혹은 [하락] 중 하나로 답하여라.

        [제목]: 장마감 이후 올라온 공시 자료를 보니, 분기 실적이 예상치를 크게 밑도네요ㅠ
        [정답]: [하락]

        [제목]: 간밤에 미국 연준의 깜짝 금리 인하 결정에 힘입어 미국 증시 폭등 중이네요. 내일 한국 증시도 기대해 봅니다!
        [정답]: [상승]


[제목]: 문제인이 북한을 옹호 하는 이유 가면가요
[정답]: [하락]<|endoftext|>
Gold Answer: [하락]
```

- Test 데이터 300건 중 148건 (49%) 정도만 정답
  - 정답인 경우에도, 객관적으로는 정답으로 보기 애매한 경우들도 많음
    - 특히, 해당 주식 종목과 직접적으로 연관되지 않은 글인 경우
- 시사점 : 개인 투자자 정서와 주가 등락 간 밀접한 관계x
  - 이전 연구(상관관계=0.03)와 대동소이한 결과

Future Works

- 데이터셋 퀄리티 개선
  - 해당 주식 종목과 무관한 정치글, 광고글 등의 비중이 생각보다 큼 (특히 대형주일수록)
    - 이러한 데이터들로 모델을 학습시키고 테스트 해 보는 것은 의미 없고 부적절
- 모델 크기 및 데이터셋 규모 확장
  - 현재까지 사용한 모델은 파라미터가 13억개로, 최신 모델들 대비 매우 적은 수준
    - LoRA, Quantization 등 모델 경량화 기법들의 도움을 받아, 보다 큰 파라미터의 모델들 활용
  - 네이버증권 외 타 도메인 (뉴스 기사 등)으로의 데이터셋 확장
    - 정제된 데이터셋 확보

Thank you!