

SECTION4 PROJECT



AIB_15 박준영

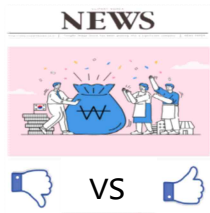
SECTION4 PROJECT

금융 뉴스 제목 감정 분석

AIB_15 박준영

Content 1	프로젝트 개요 및 문제 설정
Content 2	데이터 설명
Content 3	전처리 및 모델링
Content 4	결과 제시
Content 5	한계점, 추후 해결 방안

CONTENT 01 프로젝트 개요 및 문제 설정



- 금융상품 투자 -> 금융지표, 기업의 정량 및 정성적 정보 고려
- 금융뉴스 -> 어떤 정보? 어떠한 뉘앙스?
- 금융 투자 초보자 -> 뉴스를 보고 이해하는 것도 힘들 수 있다.
- 금융 뉴스를 자연어 처리 감정분석을 하여 투자에 도움이 되는 금융 감정 분석 프로그램을 제작을 목표

가설 : LSTM을 이용하여 금융뉴스의 긍정, 부정, 중립 감정을 판단할 수 있다

CONTENT 02 데이터 설명

labels	sentence	kor_sentence
neutral	According to Gran, the company has no plans to...	Gran에 따르면, 그 회사는 회사가 성장하고 있는 곳이지만, 모든 생산을 러시아로...
neutral	Technopolis plans to develop in stages an area...	테크노폴리스는 컴퓨터 기술과 통신 분야에서 일하는 회사들을 유치하기 위해 10만 평...
negative	The international electronic industry company ...	국제 전자산업 회사인 엘코텍은 탈린 공장에서 수십 명의 직원을 해고했으며, 이전의 ...
positive	With the new production plant the company woul...	새로운 생산공장으로 인해 회사는 예상되는 수요 증가를 충족시킬 수 있는 능력을 증가...

01

4846개

02

금융기사
영어,한국
어 문장

03

부정, 중립,
긍정

CONTENT 03 데이터 전처리 및 모델링

- 부정 : 0, 중립 : 1, 긍정 : 2
- 영어 문장 컬럼 삭제
중복치, 결측치 처리

라벨링 및 한국 기사 문장 설정

```
1 total_data['labels'] = total_data['labels'].replace(['negative', 'neutral', 'positive'], [0, 1, 2])  
2 total_data.head(3)
```

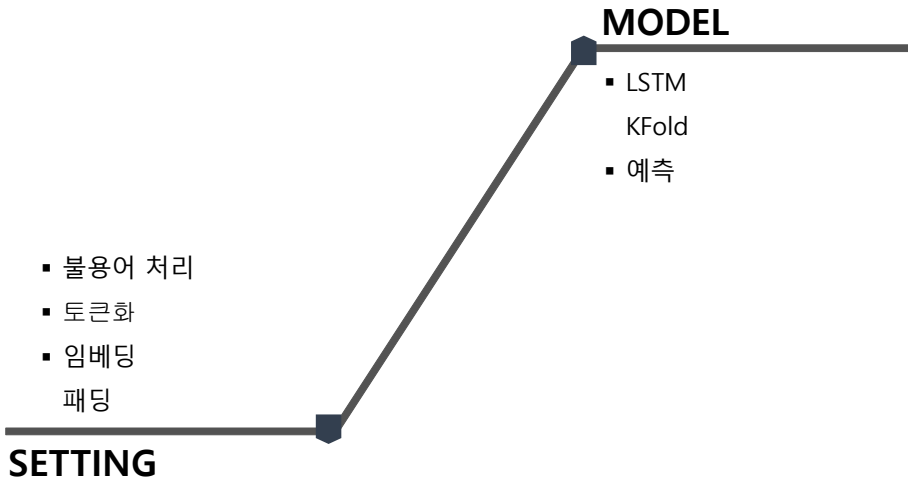
	labels	kor_sentence
0	1	Gran에 따르면, 그 회사는 회사가 성장하고 있는 곳이지만, 모든 생산을 러시아로...
1	1	테크노폴리스는 컴퓨터 기술과 통신 분야에서 일하는 회사들을 유치하기 위해 10만 평...
2	0	국제 전자산업 회사인 엘코텍은 탈린 공장에서 수십 명의 직원을 해고했으며, 이전의 ...

```
2 train_data['kor_sentence'] = train_data['kor_sentence'].str.replace("[^ㄱ-ㅎㅏ-ㅣ가-힣a-zA-Z0-9 ]", "")  
3 print(len(train_data))  
4 print(train_data[:10])
```

```
3861  
labels          kor_sentence  
4123      1      올리페카 칼라스부오가 이사회 부회장으로 선출되었다  
2676      1      Savor FACE를 Cencorp에 매각하면 받을 대가주 평가와 우세한 환율 등...  
536       1      이번 청약으로 TGK10에 대한 포럼의 지분은 주식과 의결권의 76를 약간 넘는 수...
```

- 글자와 숫자만 남기기

정규표현식 활용

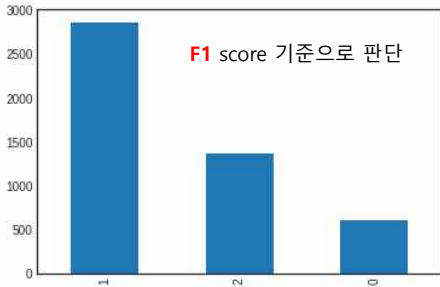


MODEL



KFold

예측



- 불용어 차
- 토큰화
- 임베딩
- 패딩

SETTING

토큰화, 임베딩

4123 [올리, 페카, 칼라스, 부오, 사회, 부회장, 으로, 선출
2676 [Savcor, FACE, Cencorp, 매각, 연, 받, 대, 가주, 평가, 우...

```
1 print(X_train[:3])  
2 print(X_test[:3])
```

```
[1086, 1355, 3088, 1537, 277, 1211, 9, 1538, 16, 22],  
[485, 55, 88, 1995, 25, 1385, 3252, 8142, 169, 35, 7,
```

```
1 print(X_train[:1])  
2 print(X_test[:1])
```

```
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  
 0  0  0  0  0  0  0 1086 1355 3088 1537 277 1211 9  
1538 16 22]  
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  
 0 485 55 88 1995 25 1385 3252 8142 169 35 7 9 654  
24 4 23]
```

패딩

LSTM

- 기울기 소실 해결
정보에 몇 % 기억
- 금융 기사 문장에 적합

```
1 model = Sequential()  
2 model.add(Embedding(vocab_size, 100))  
3 model.add(LSTM(128))  
4 # model.add(BatchNormalization())  
5 model.add(Dropout(0.5))  
6 model.add(Dense(64, activation = 'relu'))  
7 # model.add(BatchNormalization())  
8 model.add(Dropout(0.5))  
9 model.add(Dense(3, activation = 'softmax'))
```

```
model.compile(optimizer = 'adam',  
              loss = 'categorical_crossentropy',  
              metrics=[get_f1])
```

code

CONTENT 04 결과 제시

LSTM F1 : 0.7131
KFold : 0.7100

```
1 sentiment_pred('이번주 방향성 없는 코스피...시장은 재차 숨고르기')  
2 sentiment_pred('올해 마지막 FOMC 앞두고 블랙아웃 진입...피봇 기대감 여전')  
3 sentiment_pred('올해 삼성전자 임원 자사주 121억원 매입...주가부양 노력')  
4 sentiment_pred('세일즈포스, 3분기 실적 컨센 상회...계약 반등 시점은 불확실')
```

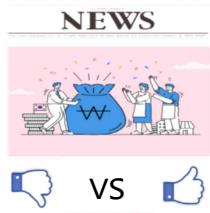
CONTENT 04 결과 제시

LSTM F1 : 0.7131
KFold : 0.7100

```
1 sentiment_pred('이번주 방향성 없는 코스피...시장은 재차 숨고르기')
2 sentiment_pred('올해 마지막 FOMC 앞두고 블랙아웃 진입...피봇 기대감 여전')
3 sentiment_pred('올해 삼성전자 임원 자사주 121억원 매입...주가부양 노력')
4 sentiment_pred('세일즈포스, 3분기 실적 컨센 상회...계약 반등 시점은 불확실')
```

```
1/1 [=====] - 0s 17ms/step
기사 제목 : 이번주 방향성 없는 코스피...시장은 재차 숨고르기 -> 중립(97.23%)
1/1 [=====] - 0s 15ms/step
기사 제목 : 올해 마지막 FOMC 앞두고 블랙아웃 진입...피봇 기대감 여전 -> 중립(97.80%)
1/1 [=====] - 0s 15ms/step
기사 제목 : 올해 삼성전자 임원 자사주 121억원 매입...주가부양 노력 -> 긍정(100.00%)
1/1 [=====] - 0s 15ms/step
기사 제목 : 세일즈포스, 3분기 실적 컨센 상회...계약 반등 시점은 불확실 -> 부정(87.30%)
```

CONTENT 05 한계점 및 추후 해결 방안



- 데이터 부족 -> 4천여개로 양적인 부족함 (+ 금융 토론방 문제)
금융 관련 문장의 다소 애매한 긍부정 단어 (감소 -> 긍정? 부정?)
- 자연어 처리 모델의 최적화 의문
- Section3 프로젝트는 실패 했는데 부족하더라도 예측 결론을 낸 프로젝트를 완성시킨 것에 대해서는 칭찬을 해주고 싶다.

크롤링을 통해 더 많은 데이터 개선 방안 고민

뉴스 기사가 아니라 기업을 입력했을때 그 기업 관련 뉴스의 감정을 분석할 수 있으면 더욱 실용성이 있을 것 같음.

감사합니다

AIB_15 박준영