BOAZ 학기 세션 딥러닝 미니 팀플 발표

감성분석을 통한 주식시장 분석

19기 분석 Sentiment Analysis 팀
송윤정, 임서현, 장우솔, 정은진

CONTENTS

1. 주제 소개

감성분석

감성 분석 정의

감성 분석

- 텍스트 분석 방법 중 하나로, 텍스트의 어조가 긍정, 부정 또는 연구자가 지정한 세부적인 감정(기쁨, 분노) 중 어느 것에 해당하는지 분류하는 방법

- Entity : 문장에서 이야기하는 주제
- Aspect : Entity의 특성
- Opinion or Sentiment : 긍/부/중립 또는 세분화된 레벨의 의견

(논문 : Aspect based Sentiment Analysis and Opinion Summarization)

적용 분야

#리뷰분석 # 설문조사 #트렌드조사 #브랜드 모니터링

BOAZ

1. 주제 소개

주식뉴스 감성분석

주제 설명

주식뉴스 감성분석

- 뉴스 데이터 혹은 소셜 데이터에 내포되어 있는 감성을 분석해서 주가를 예측

주제

- 주식뉴스 감성분석을 통해 주식시장 뉴스 심리 지수 개발
- 뉴스심리지수는 뉴스기사에 나타난 긍부정 문장을 지수화한 지표
- 경제심리 변화를 신속하게 포착하고 변동요인을 쉽게 파악하기 용이

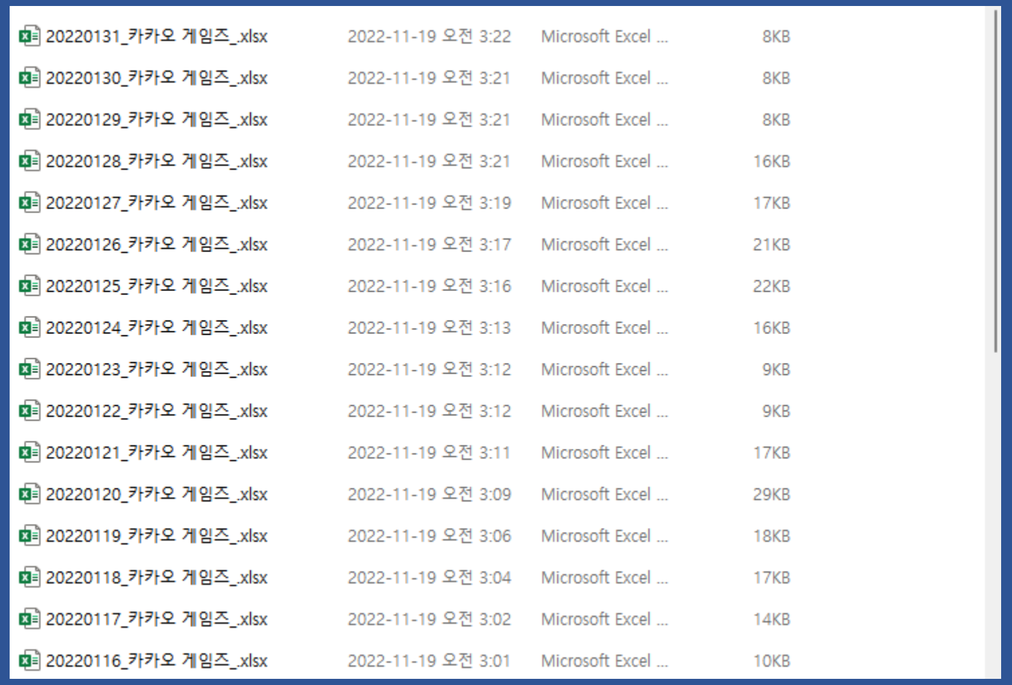
기업 선정

- 광범위한 경제 데이터 중 한 기업에 국한하기로 결정
- 게임사 중 가장 다사다난했던 **“카카오 게임즈”** 기업의 주가를 분석

BOAZ

데이터 수집과 구축

데이터 수집



20220131_카카오 게임즈_.xlsx	2022-11-19 오전 3:22	Microsoft Excel ...	8KB
20220130_카카오 게임즈_.xlsx	2022-11-19 오전 3:21	Microsoft Excel ...	8KB
20220129_카카오 게임즈_.xlsx	2022-11-19 오전 3:21	Microsoft Excel ...	8KB
20220128_카카오 게임즈_.xlsx	2022-11-19 오전 3:21	Microsoft Excel ...	16KB
20220127_카카오 게임즈_.xlsx	2022-11-19 오전 3:19	Microsoft Excel ...	17KB
20220126_카카오 게임즈_.xlsx	2022-11-19 오전 3:17	Microsoft Excel ...	21KB
20220125_카카오 게임즈_.xlsx	2022-11-19 오전 3:16	Microsoft Excel ...	22KB
20220124_카카오 게임즈_.xlsx	2022-11-19 오전 3:13	Microsoft Excel ...	16KB
20220123_카카오 게임즈_.xlsx	2022-11-19 오전 3:12	Microsoft Excel ...	9KB
20220122_카카오 게임즈_.xlsx	2022-11-19 오전 3:12	Microsoft Excel ...	9KB
20220121_카카오 게임즈_.xlsx	2022-11-19 오전 3:11	Microsoft Excel ...	17KB
20220120_카카오 게임즈_.xlsx	2022-11-19 오전 3:09	Microsoft Excel ...	29KB
20220119_카카오 게임즈_.xlsx	2022-11-19 오전 3:06	Microsoft Excel ...	18KB
20220118_카카오 게임즈_.xlsx	2022-11-19 오전 3:04	Microsoft Excel ...	17KB
20220117_카카오 게임즈_.xlsx	2022-11-19 오전 3:02	Microsoft Excel ...	14KB
20220116_카카오 게임즈_.xlsx	2022-11-19 오전 3:01	Microsoft Excel ...	10KB

데이터 설명

1. 기사 크롤링
- 네이버 뉴스 크롤링을 통해 상장한 날짜 2020년 9월 11일부터 2022년까지 약 5만 여 개의 카카오 게임즈 관련 기사 수집

2. 주가 데이터 수집
- 주가, 등락률 계산을 위한 주가 데이터 수집 (저가, 고가, 등락률 등)

데이터 라벨링

직접 라벨링 & 감성사전 구축으로 추가 라벨링

- **감성 사전 단어 예시**

✓ 긍정 단어
상승, 미소, 웃음, 탈환, 환호, 회복, 공개, 상한가, 급등, 라이온하트 상장철회, 극복, 성장, 이벤트, 투자 심리 저점 통과, 반등, 성과, 신작, 기대, 지스타, 출시, 사전예약, 유상증자, 오딘, 보라코인, 호실적, 메타버스 ,NFT, 사회공헌 캠페인, 상생, 업무협약 체결

✓ 중립단어
등락, 혼조, 공방, 보합, 복구

✓ 부정단어
신저가, 먹통, 증발, 냉담, 비판, 우마무스메, 어닝 쇼크, 악재, 바닥, 논란, 규제, 자회사 중복 상장, 블록딜, 오버행 쇼크 하락, 폭락, 급락, 축소, 위축, 위기, 한숨, 부진, 문어발

텍스트 전처리

불용어 제거, 형태소 분석, 벡터화

01 불용어 제거

```
stopwords_ko = ['!', '"', '$', '%', '&', "'", '(', ')', '*', '+', ',', '-', '.',
       '...', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', ':', '<',
       '=', '>', '?', '@', '^', '_', '`', '|', '~', '·', '–', '—', '‘',
       '’', '“', '”', '…', '、', '。', '〈', '〉', '《', '》', '가', '가까스로',
       '가령', '각', '각각', '각자', '각종', '갖고말하자면', '같다', '같이', '개의치않고', '거니와',
       '거바', '거의', '것', '것과 같이', '것들', '게다가', '게우다', '겨우', '견지에서',
       '결과에 이르다', '결국', '결론을 낼 수 있다', '겸사겸사', '고려하면', '고로', '곧', '공동으로',
       '과', '과연', '관계가 있다', '관계없이', '관련이 있다', '관하여', '관한', '관해서는', '구',
       '구체적으로', '구토하다', '그', '그들', '그때', '그래', '그래도', '그래서', '그러나', '그러니',
       '그러니까', '그러면', '그러므로', '그러한즉', '그런 까닭에', '그런데', '그런즉', '그럼',
       '그럼에도 불구하고', '그렇게 함으로써', '그렇지', '그렇지 않다면', '그렇지 않으면', '그렇지만',
```

- 불용어란?
  : 다양한 문서에서 자주 등장해 문장 의미 해석에 크게 영향을 주지 않는 단어

- 불용어 사전은 시중에 존재하는 여러 불용어 리스트를 참조하여 구축

02 형태소 분석

```
# 텍스트 정제 (형태소 분석)
for i, document in enumerate(documents):
    clean_words = []
    for word in okt.pos(document, stem=True): # 어간 추출
        if word[1] in ['Noun', 'Verb', 'Adjective']: # 명사/동사/형용사만 추출
            clean_words.append(word[0])
    document = ' '.join(clean_words)
    documents[i] = document
```

- 문장을 단어로 나누기 위해 품사를 참고
- KoNLPy의 okt 라이브러리를 활용해서 명사, 동사, 형용사를 추출

텍스트 전처리

불용어 제거, 형태소 분석, 벡터화

03 벡터화

- 벡터화란 ?
  컴퓨터가 이해할 수 있는 숫자 집합으로 텍스트를 전환하는 과정

방법 1.
카운트 기반의 단어 표현

SVM, 나이브 베이즈 모델과 같은 머신러닝 기법에 적용하기 위해
TFIDF로 전환

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf_vectorizer = TfidfVectorizer()
tf_x_data = tfidf_vectorizer.fit_transform(X) # 텍스트에 포함된 단어들의 TF-IDF 가중치 계산
```

방법 2.
워드 임베딩

FinBERT 모델을 사용하기 위해 special token과 padding 추가 후
임베딩 벡터로 전환

```
finbert_tokenizer = AutoTokenizer.from_pretrained("snunlp/KR-FinBert-SC")

# Encode the Training data
encoded_data_train = finbert_tokenizer.batch_encode_plus(
    X_train.values.tolist(),
    return_tensors='pt',
    add_special_tokens=True,
    return_attention_mask=True,
    pad_to_max_length=True,
    max_length=150 # the maximum lenght observed in the headlines
)
```

모델 설명

FinBERT

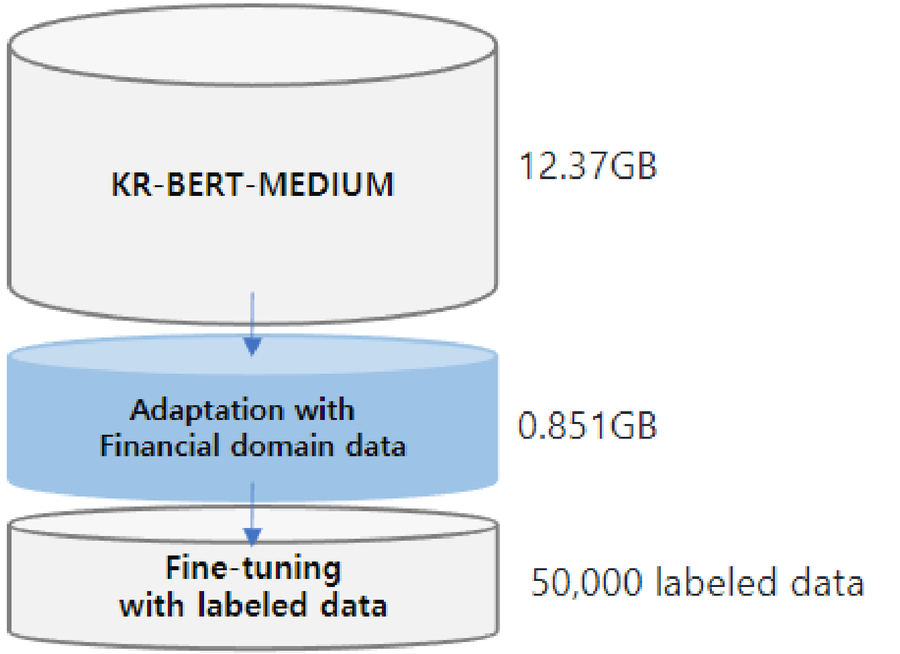
01 FinBERT

- 금융 관련 뉴스 데이터로 학습시킨 BERT 모델

- 08'~10' 발행된 Reuters TRC2 데이터 셋의 1.8백만 개 뉴스 기사로 구성된 금융 텍스트 말뭉치에 대해 사전 학습됨

- 다른 BERT와 비교하여, 금융 분야 텍스트에서 텍스트 분류 작업의 정확도가 15% 향상되어 금융 분야에서 적합함

- FinBERT를 기반으로 한국어 경제 기사를 학습한 KR-FinBERT-SC 등도 있음



3. Modeling

모델 설명

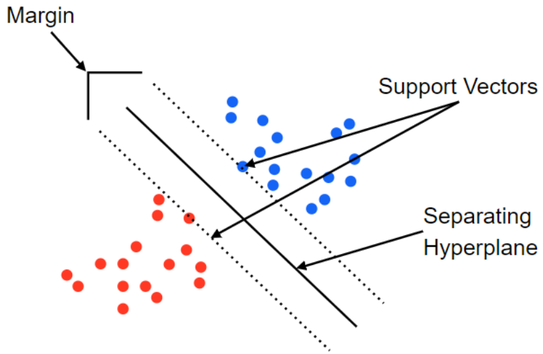
SVM, Naïve Bayes

02. SVM

- Support Vector Machine의 약자로 선형, 비선형 분류, 이상치 탐색 등 다양한 목적으로 쓰일 수 있는 머신러닝 모델

- 딥러닝에 비해 작거나 중간 사이즈의 데이터셋에 좋은 성능을 보이는 것으로 유명

- SVM은 데이터를 가장 잘 구분하는 경계선인 decision boundary를 찾는 것을 목표로 한다

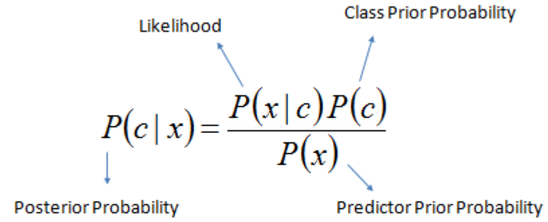


03. Naïve Bayes

- Naïve Bayes 분류 모델은 feature들 사이의 독립을 가정하는 베이즈 정리를 이용

- 조건부 독립을 가정하여 각각의 feature값의 상관관계를 확인

- 이해하기 쉽고 과정이 명료해서 많이 사용되는 지도학습 머신러닝 기법



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

4. 분석 결과

분석결과

모델 성능 결과

분석 결과

	Naïve Bayes	SVM	FinBERT
F1 Score	0.594	0.605	0.78

- F1 Score metrics를 사용해 모델별 성능을 측정
  ⇒ 레이블을 학습한 FinBERT가 가장 높았음

- 전통적인 머신러닝 기법에 비해, Transformer 모델 구조와 attention mechanism 을 활용한 딥러닝 모델의 성능이 뛰어남

BOAZ

4. 분석 결과

시각화

비교를 위한 일자별 감성 지수 산출

비지도 학습으로 라벨 구축

- FinBERT 모델로 비지도학습 방법을 사용, 긍, 부정 라벨 만듦
- 시간에 따라 직접 만든 라벨과 비지도학습으로 구축한 라벨을 시각화하여 비교
- 총 데이터는 직접 라벨링한 10,000개에서 중복 뉴스를 제외한 4,453개로 구성

```
tokenizer = AutoTokenizer.from_pretrained("snunlp/KR-FinBert-SC")
model = AutoModelForSequenceClassification.from_pretrained("snunlp/KR-FinBert-SC")
```

```
outputs_list = []
for text in tqdm(X) :
    inputs = tokenizer(text, return_tensors='pt').to(device)
    output = model(**inputs)
    output = output.logits.tolist()[0]
    outputs_list.append(output)
#출력값
outputs = torch.tensor(outputs_list)
```

```
100%|██████████| 4475/4475 [07:00<00:00, 10.65it/s]
```

```
# 확률값으로 변경
predictions=nn.softmax(outputs, -1)
#출력값 확인
df_sc = pd.DataFrame(predictions.numpy())
df_sc.columns = ['부정','중립','긍정']
df_sc
```

	부정	중립	긍정
0	0.000043	0.000129	0.999828
1	0.061545	0.737591	0.200864
2	0.001079	0.001258	0.997663
3	0.000058	0.999817	0.000125
4	0.563029	0.397754	0.039216
...	...	...	...
4470	0.000051	0.000080	0.999869
4471	0.000031	0.000220	0.999749
4472	0.095223	0.903755	0.001022
4473	0.011030	0.000840	0.988130

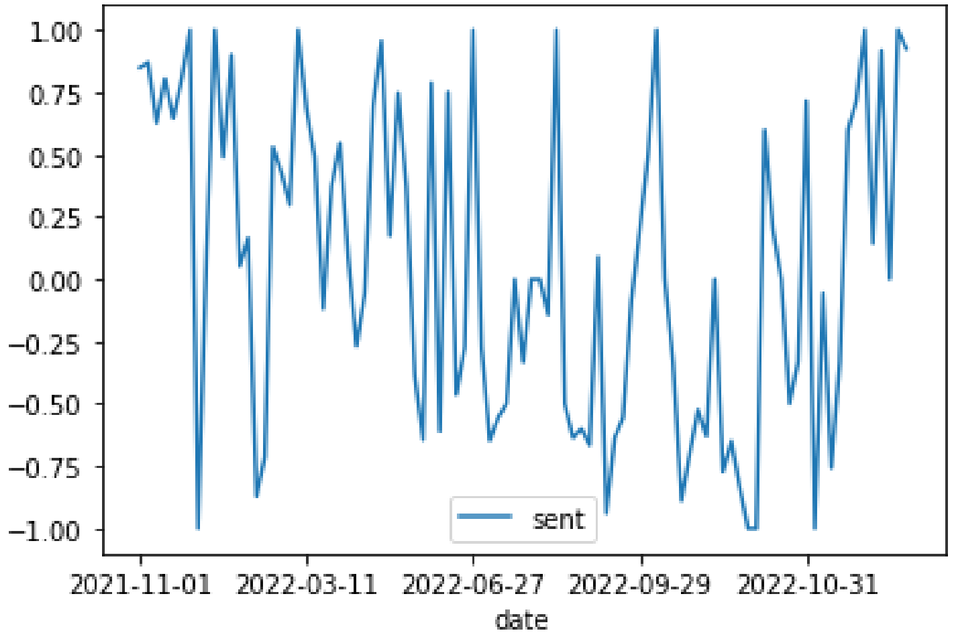
- 1) 같은 날짜에 여러 뉴스 데이터가 존재하고, 2) 날짜에 따라 뉴스 개수에 영향을 받을 수 있기 때문에
  ⇒ 일별 긍,부정 개수를 세고 -1~1사이의 값으로 표준화 진행

4. 분석 결과

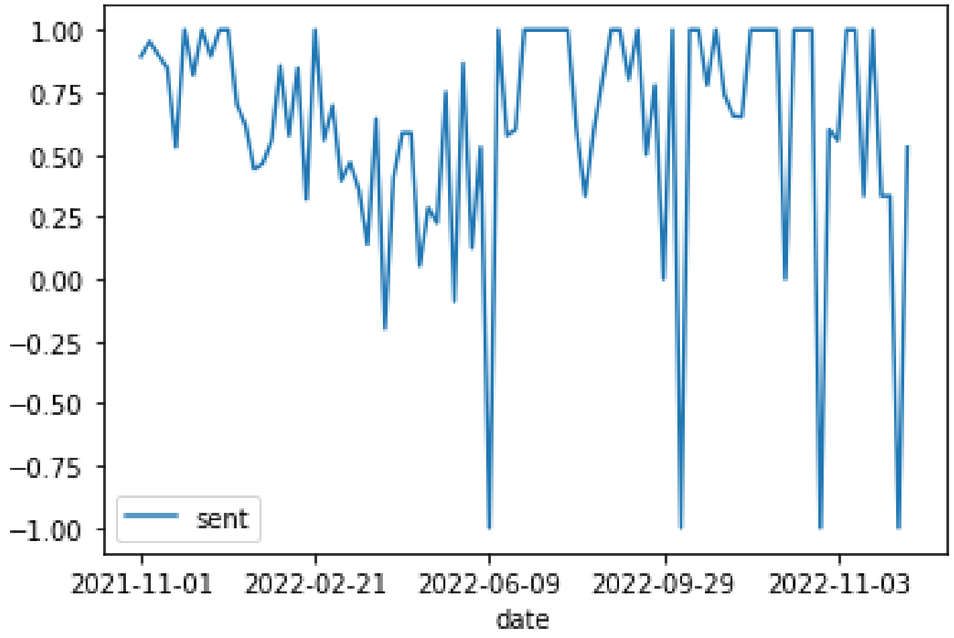
시각화

일자별 감성 지수

01 FinBert Labeling



02 직접 Labeling



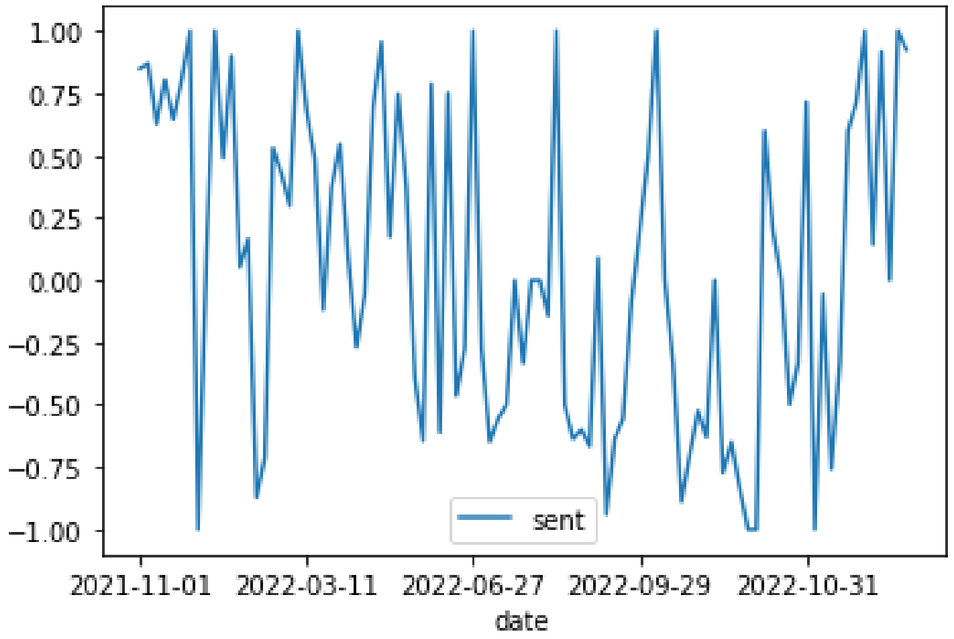
4. 분석 결과

시각화

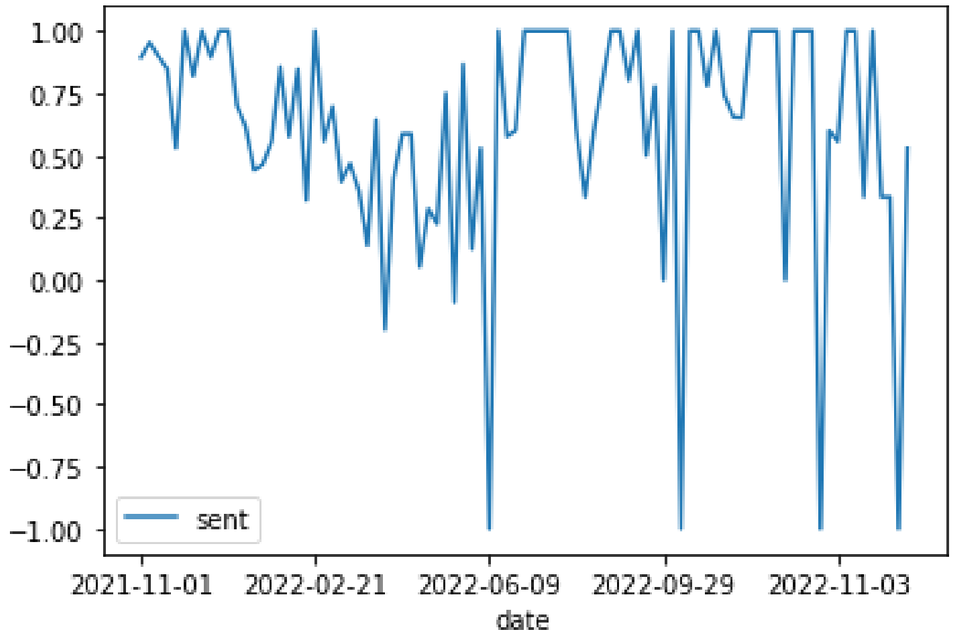
일자별 감성 지수

FinBert와 직접 Labeling 間 차이 존재

01 FinBert Labeling

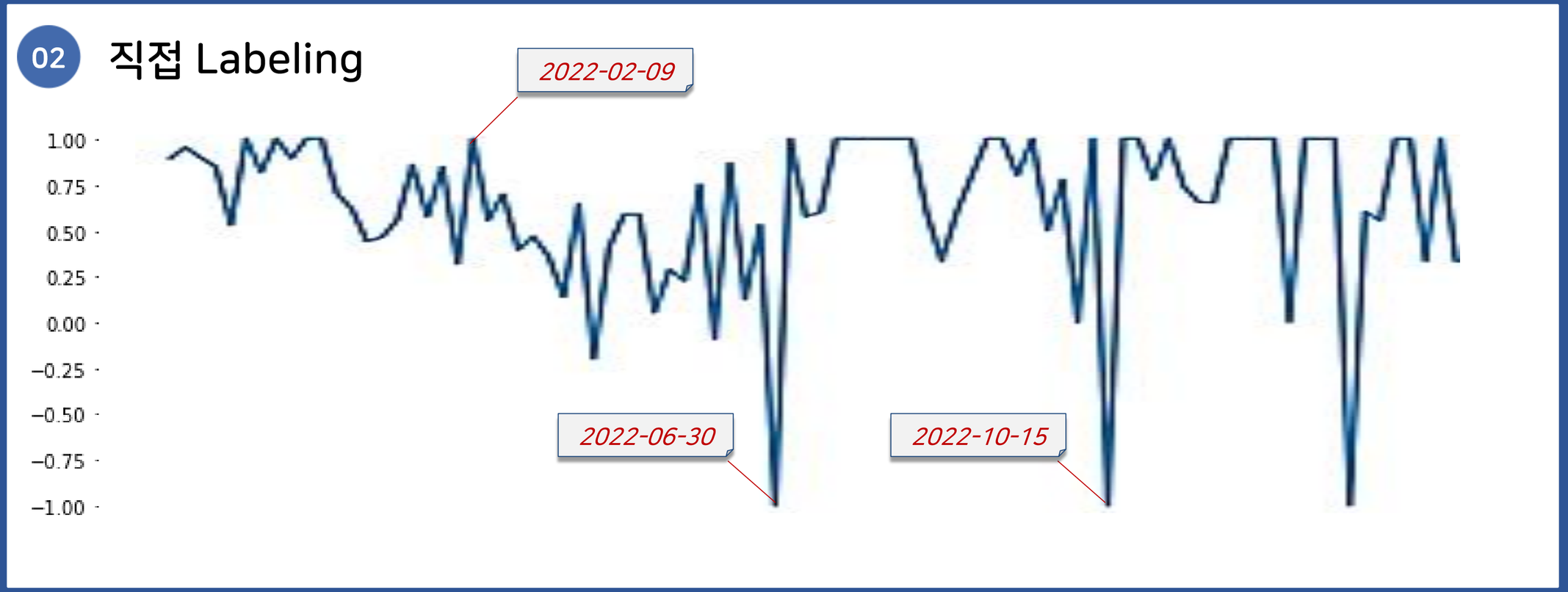


02 직접 Labeling



시각화

감성지수 추이에 따른 카카오게임즈 주요 이슈 분석



4. 분석 결과

시각화

감성지수 추이에 따른 카카오게임즈 주요 이슈 분석

02 직접 Labeling

2022-02-09

1.00

2월 9일 뉴스; 자회사 게임 '오딘' 으로 역대 최대 실적 달성

카카오게임즈, 연매출 1조 달성...올해는 'BORA 2.0' 초점(종합)

넷마블·카카오게임즈, 매출 '하이킥'

주목받는 '중견 게임사들의 반란', 카카오게임즈·위메이드 역대급 실적(종합)

'오딘으로 승승장구' 카카오게임즈, 지난해 사상 첫 연매출 1조 돌파

카카오게임즈, 2021년?매출?1조?125억 원 달성...104% 껑충

...

카카오게임즈, '오딘' 성과에 연매출 1조원 돌파

"역대 최대 실적 달성"... 카카오게임즈, 2021년 영업이익 1143억 원

'사상 최대 실적' 카카오게임즈, 연간 매출 1조125억 비결은?

[마감 시황] 외국인·기관 동반 매수에 양대지수 상승 마감

코스피 이틀째 상승 2768선 마감, 코스닥도 910선 탈환

4. 분석 결과

분석결과

워드 클라우드 시각화
Word2Vec을 활용한
연관키워드 탐색

분석 결과

1. 워드 클라우드
   - 출시 게임명 '오딘', '사상 최대', '매출 1조', '영업익', '최대 실적 달성' 등

   → 긍정적 실적 관련 단어 다수

2. '오딘' 연관 키워드
   - '성과' 98%
   - '글로벌' 99.7%
   - '흥행' 97.3%,
   - '실적', 96.5%
   - '매출' 94%

6월 30일 뉴스; 카카오뱅크의 외인·기관 공매도량 증가로 인한 코스피와 시가총액 순위 하락

[30일 마감시황] 카카오, 기아에 시총 10위 내줬다...코스피 2332.64 마감

카카오뱅크 공매도 3거래일 연속 증가...주가는 하락

[마감] 카카오뱅크, 목표주가 2만4600원 충격에 '급락'...LG엔솔 등 하락

상반기 코스피 21%·코스닥 27%↓...반년새 시총 489조원 증발

엔씨소프트, 실적감소 전망에 52주 신저가... 목표가 하향도 잇따라

[아!이뉴스] 하반기 리니지 철옹성 깨지나..."포장비도 수수료 받나요?"

코스피, 외국인·기관 대량 매도에 2,330대로 추락(종합)

코스피·코스닥, 장 초반 하락 출발

외인·기관 매도 행진...연중최저치 접근한 코스피

[시황종합] 코스피, 2330선 2분기 마감..."최악 분기"

코스피, 美 GDP·파월 발언 주시하며 하락...시총 상위株 엇갈려

다시 선 넘은 환율...코스피·코스닥 하락 출발

코스피 하락 출발, '실적 우려' 엔씨소프트 6% 급락 [개장시황]

뉴욕증시 비트코인 흔들 제롬파월 또 "자이언트스텝" PCE 물가+ GDP 성장률

4. 분석 결과

분석결과

워드 클라우드 시각화
Word2Vec을 활용한
연관키워드 탐색

분석 결과

1. 워드 클라우드
   - ‘급락’, ‘하락’, ‘시총’, ‘신저가’ 등
   → 부정적 실적 관련 단어
   - ‘카카오뱅크’, ‘외국인’, ‘매도’ 등
   → 주가 하락 원인 관련 단어
   - ‘엔씨소프트’, ‘셀트리온제약’ 등
   → 함께 코스피 하락한 연관 기업

2. ' 외인’ 연관 키워드
   - ‘하락’ 99.6%
   - ‘불안’ 97.5%
   - ‘투심’ 89.4%,
   - ‘매도’, 89.2%
   - ‘급락’ 82.3%

시각화

감성지수 추이에 대

10월 15일 뉴스; 판교 카카오 데이터 센터 화재, 카카오톡 서비스 먹통 이슈 발생

"판교 데이터 센터 화재" 카카오 주말 무더기 먹통

"폰 고장난 줄"...카카오 먹통에 전국민 '불편' 넘어 '분노' [종합]

[속보]카카오 서비스, 주말 무더기 장애..."판교 데이터센터 화재 영향"

주경제 오늘의 뉴스 종합] 데이터센터 화재로 멈춘 일상...카카오·네이버 서비스...

[종목현미경] 한고비 넘겼는데...카카오게임즈, 산 넘어 산

[종합] 데이터센터 화재로 카카오·네이버 서비스 장애...주말 인터넷 대란

[종합]카카오톡(카톡) 전송 오류 원인·복구 시간 '해결 언제'...네이버 다음 피해도

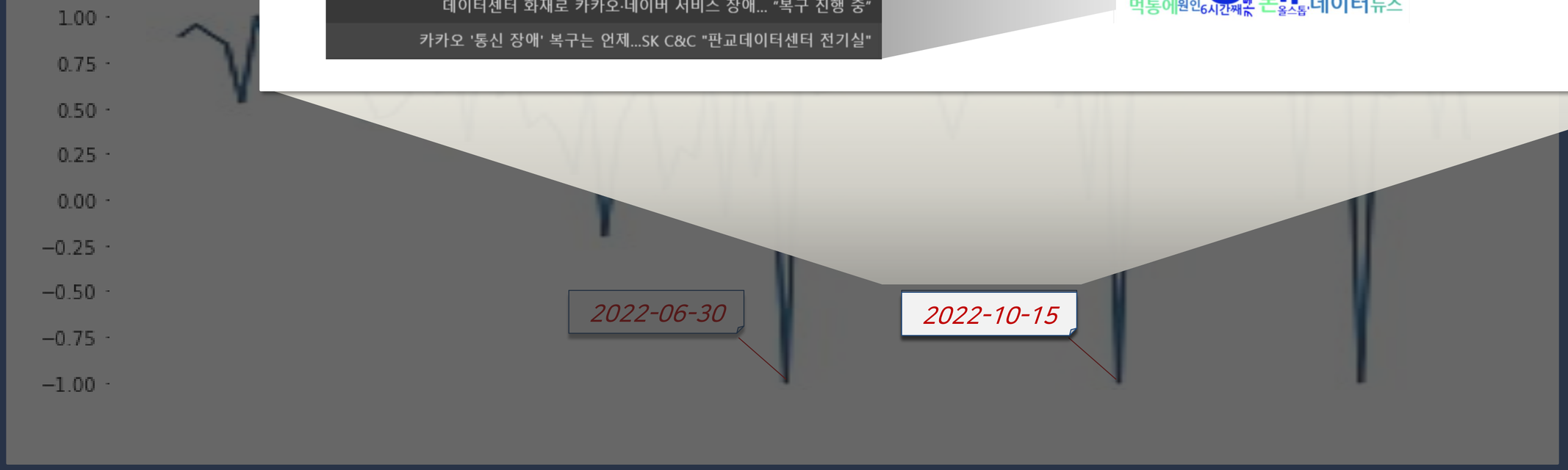
6시간째 '올스톱' 카카오...데이터센터 분리된 업무·금융만 가동중

데이터센터 화재로 카카오·네이버 서비스 무더기 장애...복구중(종합2보)

데이터센터 화재로 카카오·네이버 서비스 장애... "복구 진행 중"

카카오 '통신 장애' 복구는 언제...SK C&C "판교데이터센터 전기실"

02 직접 La



BOAZ

4. 분석 결과

분석결과

워드 클라우드 시각화,
Word2Vec을 활용한
연관키워드 탐색

분석 결과

1. 워드 클라우드
   - ‘판교‘, ‘데이터센터’, ‘무더기’, ‘먹통’, ‘화재’, ‘고장’, ‘장애‘, ‘복구‘ 등
   → 화재 사건 관련 키워드 다수
   - ‘분노’, ‘피해‘, ‘올스톱’ 등
   → 화재로 인한 심리 연관 키워드 다수

2. ' 화재’ 연관 키워드
   - ‘복구’ 89.7%
   - ‘불편’ 86%
   - ‘해결’ 56.3%,
   - ‘사퇴’ 96.8%

5. 기대 효과

기대효과 및 한계

기대효과

뉴스 감성분석 기대효과

1. 경제 상황 파악
   - 키워드 분석을 통해 신속하게 경제 상황의 변동 요인을 파악 가능

2. 모델의 자동화
   - 자료 수집에 드는 시간과 비용 절감 가능

3. 객관성
   - 모델을 통한 분석 방법으로 통계 작성자의 의중이나 실수를 차단하여 경제 상황 시에 객관성을 도모할 수 있다.

BOAZ

5. 기대 효과

기대효과 및 한계

한계 및 추가 제언

한계

- 일별 기사 개수와 라벨링 클래스가 불균형함
  ⇒ 일정한 모델 성능을 기대하기 어려움
  ⇒ Labeling 시, 빈도가 높았던 '긍정' 기사로 일괄 분류하는 경향 有

- 제목을 기준으로 분석함
  ⇒ 제목과 본문 간 다른 내용을 다룬 기사 有

추가 제언

▪ 세분화
- 더 구체적인 키워드 분석을 진행

▪ 데이터 출처의 다양화
- 뉴스 뿐만 아니라 다양한 자료를 활용
- 정확도 높은 경제 상황을 파악을 통해 더 신뢰도가 높은 자료 기대

BOAZ

[Appendix]

참고 논문

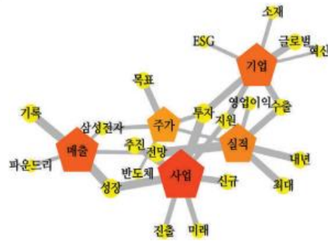
‘기계학습을 이용한 뉴스심리지수(NSI)의 작성과 활용(2022)’

서범석, 이영환, 조영배

논문 요약

- 연구 목적
  - 경제 뉴스의 NLP 분석을 통해 한국은행 공표 NSI보다 정확한 NSI 도출하고자 함
- 결론
  - 텍스트 마이닝, 기계학습과 자동화를 도입한 해당 월별 NSI는 주요 경기지표에 1~2개월 선행하는 유의미한 지표로서 사용 可

참고 내용; 분석 방법



NSI 키워드 분석

월별 NSI와 선행종합지수

코로나 발생 이후 NSI 추이

BOAZ 학기 세션 딥러닝 미니 팀플 발표

감사합니다

감성분석을 통한 주식시장 뉴스 심리지수 개발

19기 분석 송윤정, 임서현, 장우솔, 정은진