



## 주제 구체화2

날짜	@2025/03/30
다중 선택	도메인 선정 회의
활동 요약	구체적인 데이터, 주제 ,스토리라인 만들기 (텍스트 데이터 위주)

### 회의 요약

#### 영서

- 올리브영 후기의 색조화장 만족 및 불만족 요에 따른 피부타입 예측: 토픽모델과 머신러닝을 활용하여
  - 리뷰 데이터 활용, 사용자의 주관적 의견 리뷰로 만족/불만족 요인 Topic 분석
  - 토픽모델링 적용: 고객 리뷰에서 색조화장품에 대한 만족/불만족 요인
  - 머신러닝 활용 예측 모델: 만족/불만족 요인 바탕으로 고객 피부 타입(건성/지성) 예측
  - 데이터 수집 가공: 리뷰 데이터 수집 → NLP 전처리 → 토픽모델링 적용 → ML

#### 예준

- .
  - Kaggle 데이터 사용.

#### 현진

- 다이어트 식품에 대한 긍정/부정 댓글 분석
  - 제로칼로리 제품에 대해 소비자들이 기대하는 방향

### ▼ 회의 기록

### 💡 아이디어

| 텍스트 데이터를 얻기 쉬운 **트위터**를 활용한 주제를 생각해봄.

#1



### 트위터 리뷰 감성분석을 통한 브랜드 가치 (변동) 예측

(브랜드 분야: 화장품...? 구체적으로는 모르겠음)

혹은

영화 흥행 예측

- **데이터:** 특정 브랜드 광고에 작성된 인용 트윗, 혹은 특정 브랜드가 언급된 트윗 수집.
- **데이터 수집 방법:** #2의 방법들 참고.
- **연구 방법**
  - 감성분석 → 에크만의 6가지 기본 감정(anger, disgust, fear, joy, sadness, surprise)을 활용하는 분석은 대체로 딥러닝의 BERT 모델을 사용. (+회귀)
- **기타 고려사항**
  - '브랜드 가치'를 변수화하는 과정 필요 → 선행연구에서 참고할 부분이 있는지 확인.
- **관련 논문**
  - **감성분석 기반 팝업스토어 사용자 경험 분석 (2024).**
    - 국내 논문 — 정확한 피인용횟수 확인 X. (KCI에 등재되지 않은 논문인듯 함.)
    - 연구 목적: 사용자 경험에 관한 긍정/부정 요소를 분석함으로써 향후 마케팅 전략 개선과 사용자 경험 향상에 기여.
    - 연구 방법: 팝업스토어 방문 후기 관련 블로그 크롤링, 중복제거/문장 분리 등 전처리. BERT 모델을 활용하여 감성분석 수행.
  - **택배 업체의 ESG 활동에 대한 소비자 평가가 브랜드 가치에 미치는 영향 (2024).**
    - KCI 등재. KCI 피인용횟수는 0건인데(2024에 발행해서 그런 듯) 제1저자 H지수가 10.
    - '브랜드 가치' 키워드에 포함돼서 찾아둠. (분석 방법과는 관련 X)
    - 브랜드 가치 관련 선행연구들이 기재되어 있으니 참고해도 좋을 듯.
    - 연구 목적: 실증분석을 통해 택배 업체의 ESG 활동에 대한 마케팅 전략과 시사점을 제시
  - **Competitive Sentiment Analysis for Brand Reputation Monitoring (2024).**
    - Scopus FWCI 4.04
    - 연구 목적: 사람들이 브랜드에 느끼는 감정을 추적하고, 시장에서 유사 업계 브랜드에 대한 사람들의 인식과 비교하여 가치 있는 인사이트를 제공, 브랜드 평판 유지를 위한 실용적인 해결책 제언.
    - 연구 방법: 텍스트 유사도(Textual Similarity calculation)와 경쟁 분석을 위한 머신러닝 기법 활용.
      - Flask web framework> 웹 애플리케이션 개발 툴 사용
      - Sentiment Analysis> **TextBlob** 라이브러리(감성분석 단순화), **NLTK** 라이브러리(질 높은 텍스트 전처리), **VADER** 라이브러리(소셜미디어와 비정형데이터에서 감성을 측정하는 데에 사용).

## #2



## 트위터 피드 감성분석을 통한 국내 트렌드 예측 지표 제안

(ex. 리트윗/마음에 들어요 수, 인용횟수와 트렌드의 연관성 or 특정 감성과 트렌드와의 연관성)

- **데이터:** 트위터 피드의 텍스트, 바이럴 되는 트윗의 리트윗/마음에 들어요/인용 수 (수치데이터) 등.
- **데이터 수집 방법**
  - 트위터 API 활용
    - 유료 기능이 포함될 수 있음
    - 데이터 수집을 위한 라이브러리: Python의 `snsrape` 라이브러리, `tweepy` 라이브러리
  - 웹 스크래핑
    - 웹 스크래핑은 웹 페이지의 HTML 코드를 분석하여 원하는 데이터를 추출하는 방법
    - 웹 스크래핑을 위한 도구: Python의 `snsrape` 라이브러리
  - 이미 수집된 데이터 사용
    - \*Kaggle\*에서 "Twitter Sentiment Analysis" 데이터셋 검색
    - \*Hugging Face Datasets\*에서 트위터 관련 데이터 찾기
    - **GitHub**에서 공개된 트위터 데이터셋 활용
    - 가장 빠르고 간편한 방식이긴 하지만 데이터 최신성이 부족함, 원하는 방향으로 활용하지 못할 수도 있음.

## ChatGPT 검색 결과 (python 라이브러리 활용 코드 포함)

트위터(Twitter, 현재 X)의 텍스트 데이터를 수집하는 방법에는 여러 가지가 있어. 대표적인 방법을 정리해줄게.

### 1. Twitter API 사용 (공식 방법)

트위터의 공식 API를 사용하면 특정 키워드, 해시태그, 사용자 계정 등을 기준으로 데이터를 수집할 수 있어.

#### ✅ Twitter API 사용 방법

##### 1. Twitter Developer 계정 신청

- Twitter Developer Portal에 접속해 계정을 생성하고 API 사용 신청을 해야 해.
- 현재 무료(FREE) 티어에서는 기본적인 검색 기능이 제한적이므로 유료 플랜 확인 필요.

##### 2. API Key 및 Token 발급

- 계정 승인이 완료되면 API Key, API Secret Key, Access Token 등을 발급받을 수 있어.

### 3. Python 라이브러리 사용하여 데이터 수집

`tweepy` 또는 `snsrape` 라이브러리를 사용해 트위터 데이터를 가져올 수 있어.

#### 코드 예시 (Tweepy 사용)

```
import tweepy

# API 인증 정보 입력
api_key = "YOUR_API_KEY"
api_secret = "YOUR_API_SECRET"
access_token = "YOUR_ACCESS_TOKEN"
access_token_secret = "YOUR_ACCESS_TOKEN_SECRET"

# 인증 처리
auth = tweepy.OAuthHandler(api_key, api_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth, wait_on_rate_limit=True)

# 특정 키워드로 트윗 수집
query = "감성분석"
tweets = tweepy.Cursor(api.search_tweets, q=query, lang="ko", count=100).items(10)

# 출력
for tweet in tweets:
    print(tweet.text)
```

#### 주의점

- `search_tweets()` 는 과거 7일치 데이터만 제공(무료 버전).
- 트위터 정책상 `Developer 계정` 승인 과정이 필요.
- 트위터의 정책 변경으로 인해 API 사용 제한이 있을 수 있음.

### 2. SNScrape 활용 (API 없이 크롤링 가능)

`snsrape` 는 트위터 API 없이도 데이터를 수집할 수 있는 Python 라이브러리아.

## ✅ 설치 및 사용 방법

### 1. 라이브러리 설치

```
pip install snsrape
```

### 2. 특정 키워드로 트윗 수집

```
import snsrape.modules.twitter as sntwitter
import pandas as pd

query = "감성분석 since:2024-03-01 until:2024-03-30 lang:ko"
tweets = []

for tweet in sntwitter.TwitterSearchScraper(query).get_items():
    if len(tweets) >= 100: # 100개 수집 후 종료
        break
    tweets.append([tweet.date, tweet.user.username, tweet.content])

df = pd.DataFrame(tweets, columns=['Date', 'User', 'Tweet'])
print(df.head())
```

#### 📌 장점

- API 키 없이도 데이터를 수집 가능.
- 과거 데이터까지 검색 가능.

#### 📌 주의점

- 트위터의 정책 변경으로 인해 차단될 가능성이 있음.
- 너무 많은 요청을 보내면 IP가 차단될 수 있으므로 적절한 속도 조절 필요.

### 3. 트위터 데이터셋 활용 (이미 수집된 데이터 사용)

트위터에서 수집된 공개 데이터셋을 활용할 수도 있어.

- \*Kaggle\*에서 "Twitter Sentiment Analysis" 데이터셋 검색
- \*Hugging Face Datasets\*에서 트위터 관련 데이터 찾기
- **GitHub**에서 공개된 트위터 데이터셋 활용

## 📌 추천 방법

- ✅ 빠르고 간단하게 하려면? → [snsrape](#) 사용
- ✅ 공식적인 방법이 필요하다면? → [Twitter API](#) 사용
- ✅ 이미 수집된 데이터를 원한다면? → Kaggle, Hugging Face 활용

감성 분석용 트위터 데이터를 수집하려면 어떤 방식이 가장 적합할까? 😊

### • 연구 방법

1. 트렌드 측정 지표 설정.
2. 트위터의 데이터 수집, 텍스트 데이터 자연어 처리.
3. ...

### • 기타 고려사항

- 트렌드 변동 주기 / 특정 감성이 트렌드 지속 기간이나 파급력에 미치는 영향 / 트렌드가 지난 후 관심도 변동
- 트위터의 다양한 지표를 통해 트렌드 지속 기간 예측
- 어떤 것을 기준으로 트렌드 여부를 결정할 것인지 정하는 과정 필요.

### • 관련 논문

- **Harmonizing Macro-Financial and Twitter Sentiment Analysis in Forecasting Stock Market Trends (2024)**
  - Scopus에 등록되어 있지는 않음. 그냥 연구 방법 참고용으로 보면 좋을 듯.
  - 연구 목적: 트위터 감성분석을 통해 AI산업 관련 주식시장 트렌드 예측
  - 연구 방법: 자연어 처리를 통해 추출된 변수로 긍정/부정 감성분석 실행, 그래디언트 부스팅/의사결정 나무/랜덤포레스트 모델을 사용해 예측.
- **Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers (2024)**
  - Scopus FWCI 6.34
  - 연구 목적: 터키의 금융 관련 트윗들을 감성분석해 주식시장 예측.
  - 연구 방법: 사전 기반 방식과 머신러닝 기반 방식을 모두 사용.
    - 사전 기반 방식> Orange에서 제공하는 다중언어 감성 사전을 이용해 긍정/부정/중립 감성으로 라벨링 (중립 레이블은 연구에 사용되지 않음).
    - 머신러닝 기반 방식> Python의 [sklearn](#) 라이브러리를 활용, 긍정/부정으로 분류(Classification). 8:2 비율로 훈련/평가 데이터 나눔. SVM모델과 Multilayer Perceptron 분류모델 성능이 가장 good.
- **트위터 리트윗을 이용한 호감도 기반 트렌드 분석 시스템 개발 (2014)**
  - 국내 논문 — 정확한 피인용횟수 확인 X. (KCI에 등재되지 않은 논문인듯 함.)
  - 연구 목적: 리트윗 행위를 이용해 트위터 사용자가 관심을 가지는 분야, 사건 등을 분석.

- 연구 방법: N-Gram기법을 적용하여 단어 집합을 찾아 빈도수를 기준으로 트렌드 키워드를 추출, 데이터 수집은 Streaming API를 이용해 데이터를 수집(트위터의 실시간성 반영을 위함).

---

○ **토픽 모델링 기법을 이용한 트위터 트렌드 추출 (2013)**

- 국내 논문 — *정확한 피인용횟수 확인 X. (KCI에 등재되지 않은 논문인듯 함.)*
- 연구 목적: 빈도수 기반 기법과 LDA 기반 기법으로 트렌드를 추출하고 그 정확도 성능을 비교 분석.
- 연구 방법: Java 기반의 트위터 스트림 API 라이브러리인 Twitter4J를 사용하여 데이터 수집, 전처리.
  - 빈도수 기반 기법> bigram, trigram, quadgram 형태의 모든 단어 구를 추출 후 각 단어 구가 출현한 빈도수를 계산.
  - LDA 기반 기법> LDA의 오픈 소스 소프트웨어인 JGibbLDA를 사용. 확률 값이 높은 상위 10개의 단어 형태로 각 토픽을 표현 후, 각 토픽을 표현하는 단어 10개의 뜻을 조합. 해당 토픽을 단어 또는 단어 구의 형태로 레이블링(labeling)함.

---

○ **토픽 모델링을 이용한 트위터 이슈 트래킹 시스템 (2014)**

- 국내 논문 — *정확한 피인용횟수 확인 X.*
- 연구 목적: 트위터 데이터를 활용하여 트위터 상에서 어떤 이슈가 있었는지 추출하고, 이를 웹상에서 시각화 하는 트위터 이슈 트래킹 시스템 TITS(Twitter Issue Tracking System)를 설계하고 구축.
- 연구 방법: Hadoop과 MongoDB를 활용하여 분석, JSON(JavaScript Object Notation) 파일 포맷의 1억 5천만개 가량의 2013년 3월 한국어 트위터 데이터를 실험 대상으로 함. Java 프로그래밍으로 수행함.