

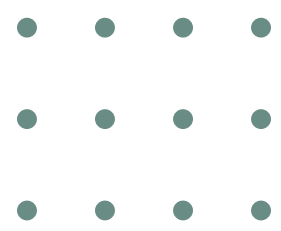


CHATGPT를 통한 감정분석

2021110977 데이터사이언스학과 조선헌
2021111336 정보보호학과 송다은
2021111727 데이터사이언스학과 엄채린

머신러닝기반데이터분석





목차

주제 선정 이유 / 문제 정의

- ChatGPT에 대한
인식을 감정분석
- ChatGPT를 통한
감정분석과 인간이
진행한 감정분석 비교

데이터 수집 및 데이터 전처리

- Kaggle ChatGPT
트위터 데이터 셋

방법론

- TF-IDF
- linearSVC
- $y = f(x)$

인사이트

- ChatGPT를 어떻게
활용해야할까

"AI, 핵처럼 규제하자"는 **챗GPT** 아버지...한국AI 숨통 조여온다 [A...
챗GPT 같은 AI 모델을 보유한 소수 미국 빅테크가 잇따라 AI 핵심 기술을 공개하
지 않은 데 이어 AI를 규제해야 한다고 주장하면서다. 오픈AI는 지난 2월 출시한 ...

| LG, 퓨리오사AI와 'AI 반도체' 동맹...**챗GP**... 헤럴드경제 | 5시간 전 | 네이버뉴스



ZDNet Korea **PICK** | 5시간 전 | 네이버뉴스

[방은주의 AI이야기] '**챗GPT** 아버지' 알트먼이 한국에 오는 이유
(현재 **챗** API는 대화 내용을 전부 보내 맥락을 기억해야 하지만 미래에는 대화 이
력을 기억하는 API가 나올 예정) 실현이 화두고, 내년에는 ▲멀티 모달리티(이미...

| '**챗GPT**' 만든 오픈AI CEO "IPO 관심 ... 조선비즈 **PICK** | 2시간 전 | 네이버뉴스



조선일보 | 7시간 전 | 네이버뉴스

팀 쿡 "**챗GPT** 다양한 프로그램에 흥분...애플도 면밀히 살피고 있...
팀 쿡 애플 최고경영자(CEO)가 6일(현지시각) 미 ABC와의 인터뷰에서 "나도 **챗GP**
T를 사용하고 있다"며 "**챗GPT**의 다양한 응용 프로그램에 대해 흥분하고 있다"고...

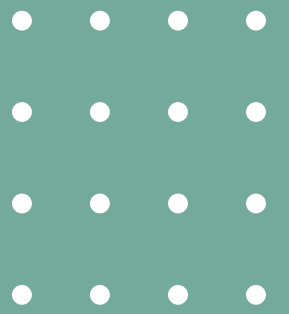
| 팀 쿡 "애플, **챗GPT** 면밀히 검토 ... ZDNet Korea **PICK** | 2시간 전 | 네이버뉴스
| 팀 쿡 애플 CEO "**챗GPT** 눈여겨보는 중" 디지털투데이 | 6시간 전



국민일보 **PICK** | 2시간 전 | 네이버뉴스

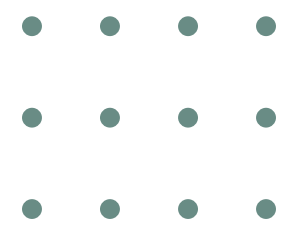


IT 시장을 바꾼 ChatGPT



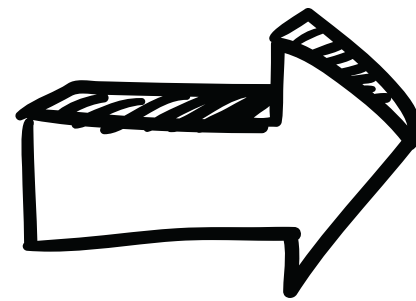
이제는
더이상
물러날 곳이 없다
ChatGPT 사용해야된다





주제 선정 이유

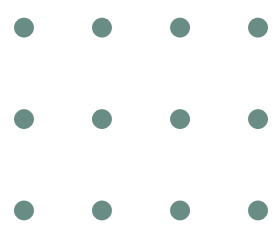
ChatGPT 사용자에 대한 인식과 관련된
데이터를 크롤링 및 감정분석을
하여 만족도 알아보기



(ChatGPT를 활용하자!!)

ChatGPT를 통해 감정분석한 모델과
인간이 만든 모델의 성능을 비교해보기



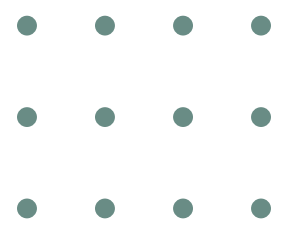


데이터 정의 & 수집

	Datetime	Text	Username	QuoteCount	ConversationId	Language	hashtag	hashtag_counts
0	2023-01-22 13:44:34+00:00	ChatGPTで遊ぶの忘れてた！！Wn書類作るコード書いてみてほしいのと、Wnどこまで思考整...	mochico0123	0	1617156270871699456	ja		0
1	2023-01-22 13:44:39+00:00	@AlexandrovnaIng Prohibition of ChatGPT has be...	Caput_LupinumSG	0	1617148639993806848	en		0
2	2023-01-22 13:44:44+00:00	Schaut Euch an, was @fobizz @DianaKnodeI alles...	ciffi	0	1617156308926349312	de	['#ChatGPT']	1
3	2023-01-22 13:44:49+00:00	Bow down to chatGPT ☐..... https://t.co/ENTSzi...	Vishwasrisiri	0	1617156332297256961	en		0
4	2023-01-22 13:44:52+00:00	Profilinde vatan, Türkiye falan yazan bireyler...	0xGenetikciniz	0	1617156345064570880	tr		0
...
49996	2023-01-24 06:57:56+00:00	#ChatGPT ist ein #Chatbot, der durch künstlich...	HorstKrieger	0	1617778712082096128	de	['#ChatGPT', '#Chatbot,', '#KI', '#AI', '#Dial...	6
49997	2023-01-24 06:57:59+00:00	@r8r Ich hab mal die AI dazu befragt (ChatGPT)...	werpu	0	1617774521586708480	de		0
49998	2023-01-24 06:58:00+00:00	5 minuti di #chatGPT e ho capito che apprende ...	marcopiccinini	0	1617778728481992705	it	['#chatGPT']	1
49999	2023-01-24 06:58:01+00:00	Portland Shop Uses ChatGPT To Tell Family Stor...	EuniceNyandat	0	1617778731678044162	en		0
50000	2023-01-24 06:58:01+00:00	Ahora sueño con el día en que Amazon integre u...	AmericoSD_69	0	1617778733355790342	es		0

- Kaggle ChatGPT Twitter Dataset 수집
(ChatGPT 관련 트윗)
- Language가 EN인 Text만 사용





크롤링 데이터

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

# Creating list to append tweet data to
# 데이터를 추가할 리스트
tweets_list2 = []

# Using TwitterSearchScrapper to scrape data and append tweets to list
# 챗GPT 라는 검색어에 대해 2023-03-14 ~ 2023-05-14 사이의 트윗을 수집한다
for i, tweet in enumerate(
    sntwitter.TwitterSearchScrapper('인어공주 OR 실사화 since:2023-04-30 until:2023-05-30').get_items()):
    # 10000개 수집하면 종료
    if i > 9999:
        break
    # 트윗의 날짜, id, 내용, 유저명 4가지를 리스트에 추가
    tweets_list2.append([tweet.date, tweet.id, tweet.content, tweet.user.username])

# Creating a dataframe from the tweets list above
# 데이터 저장된 리스트를 pandas 데이터프레임으로 변환
tweets_df2 = pd.DataFrame(tweets_list2, columns=['datetime', 'tweet_id', 'content', 'username'])

tweets_df2.to_csv('data1.csv', encoding='utf-8')
```

```
import re
import pandas as pd

# ++ ㅋㅋ 제거, ㅎㅎ 제거, 영어 제거 미완성
# Basic Cleaning Text Function
def CleanText(readData, Num=False, Eng=False):
    # Remove Retweets
    text = re.sub('RT @[w_]+: ', '', readData)

    # Remove Mentions
    text = re.sub('@[w_]+', '', text)

    # Remove or Replace URL
    text = re.sub(r"http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[*\(\)\,\:]|(?:%[0-9a-fA-F][0-9a-fA-F]))+", ' ',
        text) # http로 시작되는 url
    text = re.sub(r"[-a-zA-Z0-9@:%_\+~#={1,256}\.][a-zA-Z0-9()]{2,6}\b(?:[-a-zA-Z0-9()@:%_\+~#?&/=]*)", ' ',
        text) # http로 시작되지 않는 url

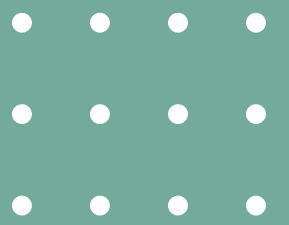
    # Remove Hashtag
    text = re.sub('[#]+[0-9a-zA-Z_]+', ' ', text)

    # Remove Garbage Words (ex. &lt; &gt; etc)
    text = re.sub('[&]+[a-z]+', ' ', text)

    # Remove Special Characters
    text = re.sub('[^0-9a-zA-Zㄱ-힣]', ' ', text)

    # Remove newline
    text = text.replace('\n', ' ')

    if Num is True:
        # Remove Numbers
        text = re.sub(r'\d+', ' ', text)
```



데이터 전처리

불필요한 열 제거

```
del df['Datetime']
del df['Username']
del df['QuoteCount']
del df['ConversationId']
del df['hashtag']
del df['hashtag_counts']
df
```

	Text	Language
0	ChatGPTで遊ぶの忘れてた！！\n書類作るコード書いてみてほしいのと、\nどこまで思考整...	ja
1	@AlexandrovnaIng Prohibition of ChatGPT has be...	en
2	Schaut Euch an, was @fobizz @DianaKnodel alles...	de
3	Bow down to chatGPT ☐..... https://t.co/ENTSzi...	en
4	Profilinde vatan, Türkiye falan yazan bireyler...	tr
...
49996	#ChatGPT ist ein #Chatbot, der durch künstlich...	de

영어만 추출

```
en = df[df['Language'] == 'en']
```

en

	Text	Language
1	@AlexandrovnaIng Prohibition of ChatGPT has be...	en
3	Bow down to chatGPT ☐..... https://t.co/ENTSzi...	en
6	ChatGPT runs 10K Nvidia training GPUs with pot...	en
7	@SWENGDAD There is repetitive work in every jo...	en
8	I created a fictional jewelry brand using Chat...	en
...
49991	I joined the @aipadtech x @moonsalecom exclusi...	en
49992	@iamjohnoliver I think ChatGPT is taking the p...	en
49993	Digital Marketers adopt new skills. Artificial...	en
49995	Remember when @Twitter was down a lot in the e...	en
49999	Portland Shop Uses ChatGPT To Tell Family Stor...	en

32076 rows × 2 columns



랜덤으로 800개의 행 추출

```
chatGPT_sentiment_english = en.sample(800, replace = False)
chatGPT_sentiment_english
```

	Text	Language
1333	Having an AMA \$TOSHI. Do not miss this one. Te...	en
40735	ChatGPT passes MBA exam given by a Wharton pro...	en
13735	ChatGPT is trending on Twitter past few weeksW...	en
32912	ChatGPT passing medical and bar exams tells yo...	en
47239	@GRDecter Let's have chat gpt do an end of lif...	en
...
11492	@David_Mulroney Maybe the Trudeau govt should ...	en
48356	@ben_golub @ChristofrayDD @AlecStapp Just copy...	en
2135	@Steve8708 Chat GPT says it's cooperative mult...	en
31494	ChatGPT's opinion on Do you think its wrong fo...	en
45243	chatGPT is insane	en

800 rows × 2 columns

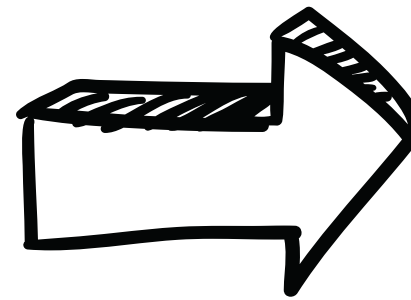

```
del chat_data['Unnamed: 0']
del chat_data['Language']
```

chat_data

Text

0	Having an AMA \$TOSHI. Do not miss this one. Te...
1	ChatGPT passes MBA exam given by a Wharton pro...
2	ChatGPT is trending on Twitter past few weeksW...
3	ChatGPT passing medical and bar exams tells yo...
4	@GRDecter Let's have chat gpt do an end of lif...
...	...
795	@David_Mulroney Maybe the Trudeau govt should ...
796	@ben_golub @ChristofrayDD @AlecStapp Just copy...
797	@Steve8708 Chat GPT says it's cooperative mult...
798	ChatGPT's opinion on Do you think its wrong fo...
799	chatGPT is insane

800 rows × 1 columns



```
real_df=pd.DataFrame(list_text, columns = ['text'])
real_df
```

text

0	having an ama toshi do not miss this one team .
1	chatgpt passes mba exam given by a wharton pro.
2	chatgpt is trending on twitter past few weeks .
3	chatgpt passing medical and bar exams tells yo.
4	let s have chat gpt do an end of life conversa.
...	...
795	maybe the trudeau govt should consult chatgpt .
796	just copy and paste a chatgpt response into a .
797	chat gpt says it s cooperative multitasking
798	chatgpt s opinion on do you think its wrong fo.
799	chatgpt is insane

800 rows × 1 columns

불필요한 열 제거(Language)

**리트윗, enticons 제거, URL 제거,
Hashtag제거, 쓰레기 단어 제거,
특수문자 제거, 피어쓰기 제거,
소문자/대문자 구분 x,**

데이터 라벨링 – ChatGPT

긍정/부정/중립으로 구분

JU

0 having an ama toshi do not miss this one team building constantly and i know this will go some decent rounds and could make a change in the space 4pm utc and their tools are just some of the topics

1 chatgpt passes mba exam given by a wharton professor

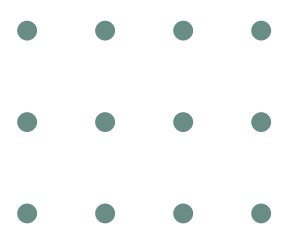
2 chatgpt is trending on twitter past few weeks here are 5 insightful threads about chatgpt you need to know

3 chatgpt passing medical and bar exams tells you more about exams than about ai

0번부터 40번까지의 문장들이 긍정인지 부정인지 중립인지 각 문장별로 구분해서 순서대로 나열해줘

문장 0: 중립 (Neutral)

문장 1: 긍정 (Positive)

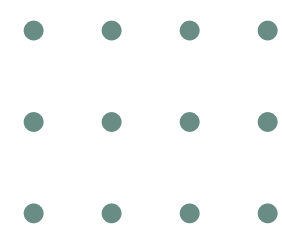


데이터 전처리 – ChatGPT

	A	B
1	0	중립
2	1	긍정
3	2	긍정
4	3	부정
5	4	중립
6	5	중립
7	6	중립
8	7	중립
9	8	중립
10	9	긍정
11	10	긍정

12	11	긍정
13	12	긍정
14	13	중립
15	14	긍정
16	15	긍정
17	16	긍정
18	17	긍정
19	18	부정
20	19	중립
21	20	중립

- ChatGPT가 판단한 800개의 감정분석 결과를 엑셀파일에 정리
- 중립 → 부정으로 값 수정



분석 결과 – ChatGPT

```
# 문장과 라벨 분리
sentences = data["문장"]
labels = data["감정"]

# train, test 데이터셋 분리
train_sentences, test_sentences, train_labels, test_labels = train_test_split(sentences, labels, test_size=0.2, random_state=42)

# TF-IDF 벡터화
vectorizer = TfidfVectorizer()
train_features = vectorizer.fit_transform(train_sentences)
test_features = vectorizer.transform(test_sentences)

# 분류 모델 학습
model = LinearSVC()
model.fit(train_features, train_labels)

# 테스트 데이터셋 예측
predictions = model.predict(test_features)

# 분류 결과 출력
print(classification_report(test_labels, predictions))
```

	precision	recall	f1-score	support
긍정	0.43	0.08	0.14	36
부정	0.78	0.97	0.87	124
accuracy			0.77	160
macro avg	0.61	0.53	0.50	160
weighted avg	0.70	0.77	0.70	160

precision
긍정/부정이라고 예측한 데이터 중
진짜 긍정/부정인 비율
긍정 : 0.43
부정 : 0.78

recall
실제 긍정인 데이터(TP+FN)들 중
긍정이라고 예측한 비율
0.08
실제 부정인 데이터(TP+FN)들 중
부정이라고 예측한 비율
0.97

정확도 : 0.77 → 77%



데이터 전처리 및 분석 - 인간

```
result_tokens = []

for i in text:
    word_tokens = i
    stop_words = set(stopwords.words('english'))

    word_tokens = word_tokenize(i)
    result = []

    for w in word_tokens:
        if w not in stop_words:
            result.append(w)
    result = ' '.join(result)

    result_tokens.append(result)

result_tokens
```

```
text = Twitter_use_only(df)

stop_words = set(stopwords.words('english'))
tokenizer = TreebankWordTokenizer()

100%|██████████| 800/800 [00:00<00:00, 20265.03it/s]
```

**ChatGPT와는 달리 인간이 하는 감정분석 과정에서는
단어로 나눠서(토큰화) → 감정을 분석**

데이터 분석 결과 - 인간

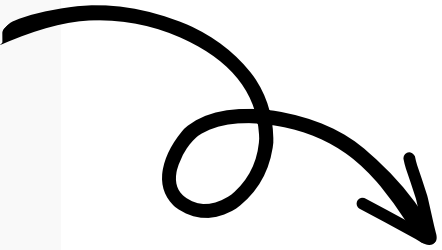
```
import nltk
from nltk.sentiment import SentimentIntensityAnalyzer

sia = SentimentIntensityAnalyzer()

n_score = 0
p_score = 0
ne_score = 0

for idx, i in enumerate(result_tokens):
    score = sia.polarity_scores(i)
    print(i)
    print(score)
    if score['compound'] >= 0.1:
        sentiment = '긍정'
        emo_df['emotion'][idx] = 1
        p_score+=1
    elif score['compound'] <= -0.1:
        sentiment = '부정'
        emo_df['emotion'][idx] = 0
        n_score+=1
    else:
        sentiment = '중립'
        emo_df['emotion'][idx] = 0
        ne_score+=1

print("긍정 문장 개수 : {}".format(p_score))
print("부정 문장 개수 : {}".format(n_score))
print("중립 문장 개수 : {}".format(ne_score))
```



ama toshi miss one team building constantly know go decent rounds could make change space 4pm utc tools topics
{'neg': 0.082, 'neu': 0.918, 'pos': 0.0, 'compound': -0.1531}
chatgpt passes mba exam given wharton professor
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
chatgpt trending twitter past weeks 5 insightful threads chatgpt need know
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
chatgpt passing medical bar exams tells exams ai
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
let chat gpt end life conversation patient family worried
{'neg': 0.216, 'neu': 0.784, 'pos': 0.0, 'compound': -0.296}
asked chatgpt scott pro vax
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
chatgpt 3 years
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
using chatgpt similar instead google searches little alpha

긍정 문장 개수 : 359
부정 문장 개수 : 122
중립 문장 개수 : 319


```

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import LinearSVC
from sklearn.metrics import classification_report

# 문장과 라벨 분리
sentences = emo_df["text"]
labels = emo_df["emotion"]

# train, test 데이터셋 분리
train_sentences, test_sentences, train_labels, test_labels = train_test_split(sentences, labels, test_size=0.2, random_state=42)

# TF-IDF 벡터화
vectorizer = TfidfVectorizer()
train_features = vectorizer.fit_transform(train_sentences)
test_features = vectorizer.transform(test_sentences)

# 분류 모델 학습
model = LinearSVC()
model.fit(train_features, train_labels)

# 테스트 데이터셋 예측
predictions = model.predict(test_features)

# 분류 결과 출력
print(classification_report(test_labels, predictions))

```

	precision	recall	f1-score	support
0	0.74	0.85	0.79	85
1	0.79	0.67	0.72	75
accuracy			0.76	160
macro avg	0.77	0.76	0.76	160
weighted avg	0.77	0.76	0.76	160

precision

긍정/부정이라고 예측한 데이터 중
진짜 긍정/부정인 비율

긍정 : 0.79
부정 : 0.74

recall

실제 긍정인 데이터(TP+FN)들 중
긍정이라고 예측한 비율

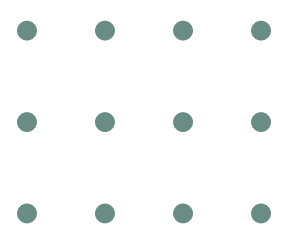
0.67

실제 부정인 데이터(TP+FN)들 중
부정이라고 예측한 비율

0.85

정확도 : 0.76 → 76%





$$Y=f(X)$$

X= 토큰화된 단어 데이터

Y= 긍정/부정을 라벨링한 데이터(0, 1)

```
[421] print(y)
```

```
0      0
1      0
2      0
3      0
4      0
```

```
..
795    0
796    0
797    0
798    0
799    0
```

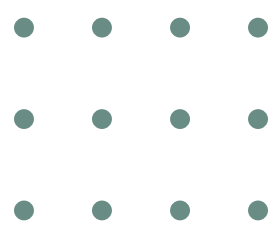
```
Name: emotion, Length: 800, dtype: int64
```



```
print(X)
```



```
(0, 2955)    0.24193415934700888
(0, 2952)    0.19117227221632843
(0, 3076)    0.26855440030240646
(0, 57)       0.26855440030240646
(0, 2707)    0.2204421767853092
(0, 511)     0.2204421767853092
(0, 1798)    0.15863189141830028
(0, 690)     0.16720169487451397
(0, 2516)    0.26855440030240646
(0, 773)     0.26855440030240646
(0, 1280)    0.19382193582991158
(0, 1641)    0.17155407052378627
(0, 649)     0.23336435589079518
(0, 438)     0.2529825575833374
(0, 2873)    0.23336435589079518
(0, 2037)    0.16585401661076585
(0, 1884)    0.23336435589079518
(0, 2957)    0.26855440030240646
(0, 186)     0.26855440030240646
```



$$Y=f(X)$$

```
# TF-IDF 벡터화
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(x)
X_train = vectorizer.fit_transform(x_train)
X_test = vectorizer.transform(x_test)

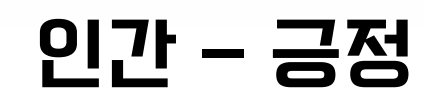
# 분류 모델 학습
model = LinearSVC()
model.fit(X_train, y_train)
```

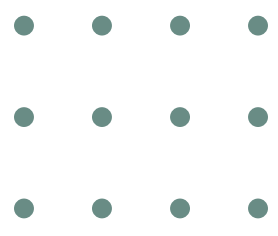
```
clf = SVC(C=10, kernel='linear', random_state=100) # SVC 클래스 생성
clf.fit(X, y) ## 모형 학습

## 파라미터 추정치
print('가중치 :', clf.coef_) # 가중치는 kernel = 'linear'인 경우만 존재
```

```
가중치 : (0, 2874) 0.0822990879508975
(0, 2819) 1.0439681992504493
(0, 426) 0.4758914335906621
(0, 2029) 0.7274370631343496
(0, 766) 0.7274370631343496
(0, 2792) 0.43386762623315817
(0, 2515) 0.43386762623315817
(0, 1945) 0.8677352524663163
(0, 1505) 0.43386762623315817
(0, 983) 0.43386762623315817
(0, 943) 0.43386762623315817
(0, 879) 0.43386762623315817
(0, 2637) 0.01794346546795667
(0, 1853) 0.03588693093591334
(0, 990) 0.01794346546795667
(0, 648) 0.01794346546795667
(0, 3084) 1.0450272867733421
(0, 2664) 1.0450272867733421
(0, 2847) 0.019165317727357824
(0, 1638) 0.019165317727357824
(0, 1421) 0.019165317727357824
(0, 1338) 0.019165317727357824
(0, 1325) 0.019165317727357824
(0, 1277) 0.019165317727357824
(0, 913) 0.019165317727357824
```

$$y = 0.082 x_1 + 1.043 x_2 + 0.475 x_3 + 0.727 x_4 + 0.727 x_5 \dots$$





결론 1

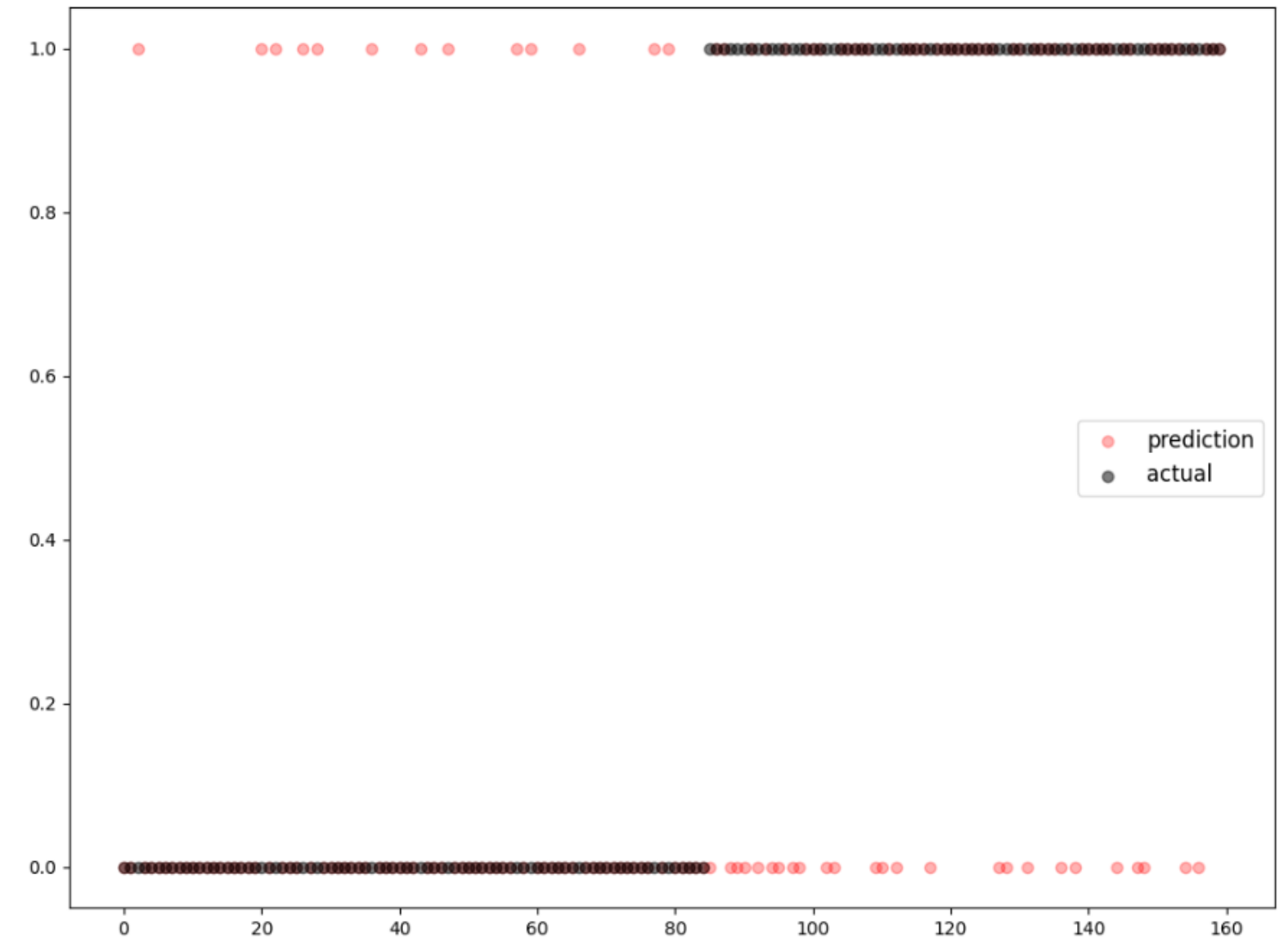
ChatGPT

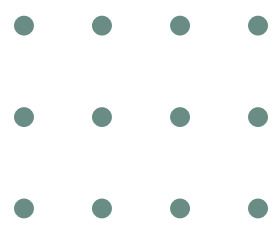
	precision	recall	f1-score	support
긍정	0.43	0.08	0.14	36
부정	0.78	0.97	0.87	124
accuracy			0.77	160
macro avg	0.61	0.53	0.50	160
weighted avg	0.70	0.77	0.70	160

인간

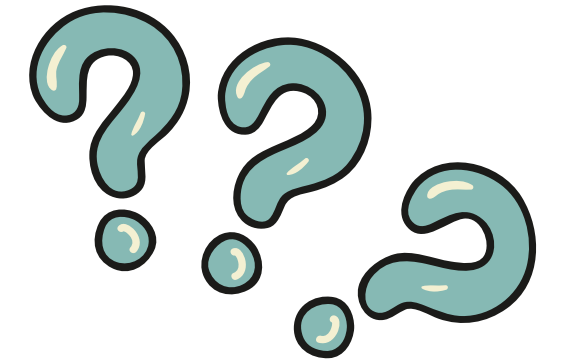
	precision	recall	f1-score	support
0	0.74	0.85	0.79	85
1	0.79	0.67	0.72	75
accuracy			0.76	160
macro avg	0.77	0.76	0.76	160
weighted avg	0.77	0.76	0.76	160

Prediction Result in Test Set





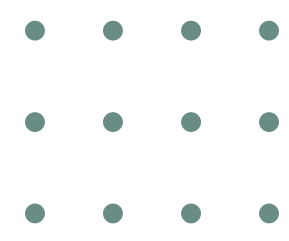
결론 2



	precision	recall	f1-score	support
긍정	0.43	0.08	0.14	36
부정	0.78	0.97	0.87	124
accuracy			0.77	160
macro avg	0.61	0.53	0.50	160
weighted avg	0.70	0.77	0.70	160

recall 값이 0.08로 정말 낮다!!
why??

- 1) 문장을 주었을 때, ChatGPT의 감정부여에서
중립 비율이 너무 높았음**
- 2) 중립을 부정에 포함시켜서 부정 비율이 너무 높음**



챗GPT로 긍정/부정(중립X) 라벨링 실패!!



아래는 주어진 문장들을 긍정, 부정으로 구분하여 순서대로 정리한 결과입니다:

500: 긍정
505: 부정
507: 부정
508: 부정
509: 긍정
510: 긍정
511: 긍정
512: 중립
513: 중립
514: 긍정
515: 중립
516: 긍정
517: 중립
518: 긍정

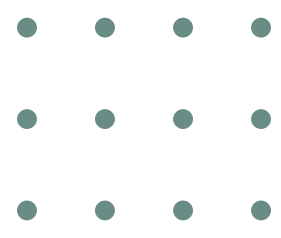
주의: 이 판단은 주관적이며, 각 문장의 문맥을 완전히 이해하지 못하고 있는 상황에서 이루어졌습니다. 따라서 결과가 정확하지 않을 수 있습니다. 전체 문장의 내용과 문맥을 파악하는 것이 중요하며, 각 문장의 긍정 또는 부정적인 의미를 파악하기 위해서는 추가 정보가 필요합니다.



문장들의 긍정 또는 부정 여부는 주어진 정보만으로는 명확히 판단할 수 없습니다. 몇 가지 문장은 긍정적인 표현을 포함하고 있지만, 다른 문장들은 제한된 정보로는 긍정 또는 부정을 판단하기 어려운 내용을 다루고 있습니다. 또한, 문장들 사이에는 명확한 순서가 없습니다. 추가적인 정보가 주어진다면 더 정확한 판단을 내릴 수 있을 것입니다.

recall 값을 높여주기 위해 ChatGPT한테
중립을 제외하고 긍정, 부정으로만 감정분석을 해달라고 함

그러나 ChatGPT는 중립을 포함해서 주거나,
명확히 판단할 수 없다는 이유로 분석을 하지 않음



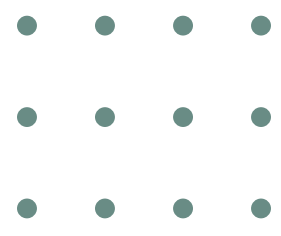
인사이트

정확도는 ChatGPT가 **77%**, 인간이 **76%**로 거의 비슷함

ChatGPT 를 활용하는 것이 **시간**과 **정확도** 부분에서 좋음

그러나 ChatGPT 사용시 감정부여에 있어서 문장을 대부분 **중립**으로 부여하는 경향이 있으므로 주의

1. 중립을 제외할 수 있도록 ChatGPT를 유도하는 **질문**을 잘 만드는 것을 추천
2. 라벨링이 긍정, 부정으로 **명확하게** 나뉘어질 수 있는 데이터를 추천



ChatGPT의 활용



머신러닝 발표 주제로 이 코드를 발표하는 것은 어떻게 생각해?



트위터 데이터셋을 활용한 감정 분석과 SVM 모델을 활용한 분류에 관한 발표는 좋은 주제입니다. 이 코드를 발표로 사용하는 것은 다음과 같은 이점을 가질 수 있습니다:

1. 실전 예제: 코드를 사용하여 실제 데이터셋에 대한 감정 분석과 분류를 수행하는 것은 실전적인 예제로서 학습자들에게 매력적일 수 있습니다. 이를 통해 머신러닝과 자연어 처리 분야에서의 실제 문제에 대한 이해를 돕고, 적용 가능성을 보여줄 수 있습니다.
2. 데이터 전처리와 텍스트 처리: 발표에서 코드의 전처리 과정과 텍스트 처리 기술에 대해 설명하고 시연함으로써, 데이터 전처리와 텍스트 처리의 중요성과 기법에 대한 이해를 제공할 수 있습니다.
3. 감정 분석: 코드에서 사용되는 감정 분석 알고리즘과 SentimentIntensityAnalyzer를 소개하고, 감정 점수 계산 및 분류 결과에 대한 설명을 통해 자연어 처리에서의 감정 분석의 원리와 활용 가능성을 보여줄 수 있습니다.
4. SVM 분류 모델: 코드에서 SVM(Support Vector Machine) 모델을 사용하여 분류를 수행하는 방법을 설명하고, 분류 결과 및 성능 평가에 대해 다루어 SVM의 개념과 활용에 대한 이해를 돕습니다.
5. 시각화: 코드에서 생성된 워드 클라우드를 활용하여 분류 결과를 시각적으로 표현함으로써, 데이터 시각화의 중요성과 텍스트 데이터에서의 시각화 기법을 소개할 수 있습니다.
6. 발표 소스 코드: 코드를 발표에서 실제로 사용하는 것은 학습자들에게 코드의 동작과 원리를 직접 보여주고 이해를 돕는 데 도움이 될 수 있습니다.

이러한 이유로, 트위터 데이터셋을 감정 분석하고 SVM 모델을 활용한 분류에 관한 발표는 흥미로울 것이며, 머신러닝과 자연어 처리에 관심이 있는 학습자들에게 유익한 내용일 것입니다.

챗gpt 관련 트위터 데이터셋을 감정분석한 코드인데 어때?



주어진 코드는 트위터 데이터셋을 전처리하고, 감정 분석을 수행한 뒤, SVM(Support Vector Machine)을 사용하여 분류 모델을 학습하고 예측하는 과정을 포함하고 있습니다. 다음은 코드의 주요 기능과 사용된 라이브러리에 대한 간략한 설명입니다:

1. 데이터 전처리: 주어진 트위터 데이터셋에 대해 기본적인 텍스트 전처리를 수행합니다. 이 과정에서 리트윗(Retweets), 멘션(@), URL, 해시태그(#), 특수문자 등을 제거하고, 불용어(stopwords)를 제거합니다.
2. 감정 분석: NLTK의 SentimentIntensityAnalyzer를 사용하여 각 문장의 감정 점수를 계산합니다. 점수에 따라 긍정, 부정, 중립으로 분류합니다.
3. 워드 클라우드: 긍정 및 부정 문장에 대해 각각 워드 클라우드를 생성하여 단어의 빈도를 시각화합니다.
4. SVM 모델 학습 및 예측: 데이터를 TF-IDF 벡터화하고, LinearSVC를 사용하여 감정 분류 모델을 학습합니다. 학습된 모델을 사용하여 테스트 데이터셋에 대한 예측을 수행하고, 분류 결과를 출력합니다.

코드 자체는 트위터 데이터셋을 감정 분석하는 데 활용할 수 있을 것입니다. 그러나 실제로 코드를 실행하려면 코드에서 참조하는 데이터 파일인 "data2.csv"를 적절한 형식으로 제공해야 합니다. 또한, 사용되는 라이브러리(nltk, pandas, sklearn)가 설치되어 있어야 합니다.

코드의 결과 및 성능은 데이터셋과 사용되는 모델에 따라 다를 수 있으므로, 실제 데이터셋과 도메인에 맞게 적용하고 결과를 평가해야 합니다.



감사합니다.