컨텐츠 주제와 타겟 국가 선정에 따른 성장가능성 예측

컴퓨터공학과 2022108115
이우석

01 서론 / 필요성

소셜 미디어의 영향력

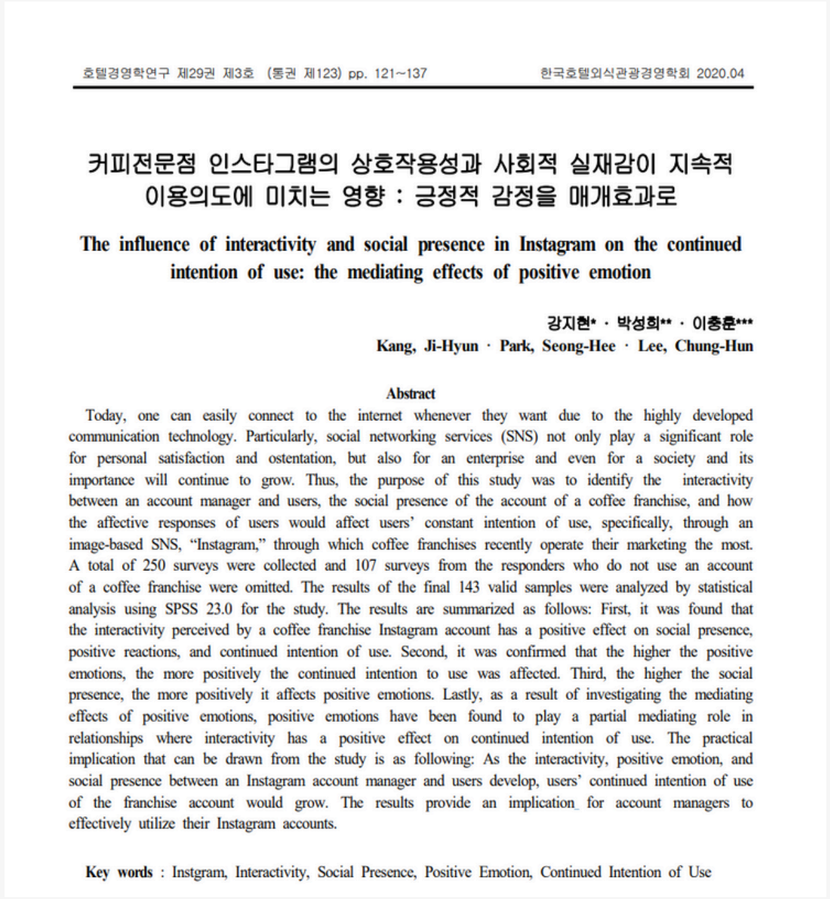
소셜 미디어, 특히 **인스타그램**은 소비자 행동과 글로벌 여론 형성에 강력한 영향을 미칩니다.
사용자의 참여를 유도하고 브랜드 인지도 및 충성도를 높이는 데 중요한 역할을 하며,
시각적 콘텐츠의 중요성이 증가하면서 **마케팅 전략의 핵심 채널**로 자리잡고 있습니다.
이러한 플랫폼은 제품 및 서비스의 홍보를 통해 소비자와의 직접적인 연결을 강화하고 있습니다.

01 서론 / 필요성

연구 목표

이 연구의 목표는 인스타그램의 컨텐츠 주제와 타겟 국가에 따른 성장 가능성을 예측하는 것입니다.
이를 통해 소비자 참여와 성과의 관계를 분석하고, 효과적인 마케팅 전략 수립을 지원하는 데 기여하고자 합니다.
또한, 국가별 소비자 특성과 선호 주제를 파악하여 인스타그램 **마케팅의 효율성을 극대화**하는 방법을 모색합니다.

02 관련 연구 / 내용



호텔경영학연구 제29권 제3호 (통권 제123) pp. 121~137 한국호텔외식관광경영학회 2020.04

커피전문점 인스타그램의 상호작용성과 사회적 실재감이 지속적 이용의도에 미치는 영향 : 긍정적 감정을 매개효과로

The influence of interactivity and social presence in Instagram on the continued intention of use: the mediating effects of positive emotion

강지현* · 박성희** · 이충훈***
Kang, Ji-Hyun · Park, Seong-Hee · Lee, Chung-Hun

Abstract

Today, one can easily connect to the internet whenever they want due to the highly developed communication technology. Particularly, social networking services (SNS) not only play a significant role for personal satisfaction and ostentation, but also for an enterprise and even for a society and its importance will continue to grow. Thus, the purpose of this study was to identify the interactivity between an account manager and users, the social presence of the account of a coffee franchise, and how the affective responses of users would affect users' constant intention of use, specifically, through an image-based SNS, "Instagram," through which coffee franchises recently operate their marketing the most. A total of 250 surveys were collected and 107 surveys from the responders who do not use an account of a coffee franchise were omitted. The results of the final 143 valid samples were analyzed by statistical analysis using SPSS 23.0 for the study. The results are summarized as follows: First, it was found that the interactivity perceived by a coffee franchise Instagram account has a positive effect on social presence, positive reactions, and continued intention of use. Second, it was confirmed that the higher the positive emotions, the more positively the continued intention to use was affected. Third, the higher the social presence, the more positively it affects positive emotions. Lastly, as a result of investigating the mediating effects of positive emotions, positive emotions have been found to play a partial mediating role in relationships where interactivity has a positive effect on continued intention of use. The practical implication that can be drawn from the study is as following: As the interactivity, positive emotion, and social presence between an Instagram account manager and users develop, users' continued intention of use of the franchise account would grow. The results provide an implication for account managers to effectively utilize their Instagram accounts.

Key words : Instgram, Interactivity, Social Presence, Positive Emotion, Continued Intention of Use

커피전문점 인스타그램의 상호작용성과 사회적 실재감이 지속적 이용의도에 미치는 영향

해당 논문에서는 마케팅 수단인 인스타그램 커피전문점 계정을 대상으로 연구를 진행하여 상호작용성, 사회적 실재감, 긍정적 감정, 지속적 이용의도의 인과관계를 실증하였다.

인스타그램 계정을 통한 마케팅이 사용자의 사회적 실재감, 긍정적 감정, 지속적 이용의도에 각각 긍정적인 영향을 미치는 것으로 나타났으며 이를 통해 SNS의 전략적 활용 등 실무적 시사점을 확인할 수 있다.

03 제안 내용

소셜 미디어 마켓팅 전략 최적화를 목표로
인스타그램 인플루언서들의 컨텐츠 주제와, 해당 인플루언서의 타겟 국가를 분석하여 어떤 주제와 타겟 국가를 선정하는 것이 효율적인 마케팅으로 이어질 수 있는지 분석하는 프로그램을 만들고자 합니다.

03-1 프로젝트 진행과정

01

데이터 수집 및 전처리

인스타그램 데이터를 로드하고 결측값 처리, 범주형 데이터 인코딩, 수치형 데이터 스케일링 등을 통해 학습에 적합한 데이터셋을 만듭니다.

02

모델 학습 및 선택

랜덤 포레스트, XGBoost, **선형 회귀** 등 다양한 머신러닝 모델을 사용하여 성장 가능성을 예측하고, 성능이 가장 좋은 모델을 최종 선택합니다.

03

결과 해석 및 평가

모델의 예측 성능을 평가하여 주요 지표(RMSE, R^2 등)를 기반으로 전략적 인사이트를 도출합니다.

해당 인사이트를 바탕으로 주제와 타겟 국가에 대한 성장 가능성을 제안하여 효과적인 마케팅 방향을 제시합니다.

04 데이터 수집 및 전처리

데이터는 kaggle의 데이터셋을 이용

Data Analysis on Top Instagram Popular Influencers

- https://www.kaggle.com/datasets/ankulsharma150/marketing-analytics-project
- 해당 데이터셋에서는 특정 컨텐츠의 주제가 많이 소비 될 것으로 예상되는 지역과 실제 데이터의 차이를 이야기함.
- 인스타그램 계정 중 팔로워 수가 가장 많은 컨텐츠는 구기종목. 하지만 축구와 관련이 깊은 영국에서의 팔로우 비중은 상위권에 속하지 못함
- 따라서 컨텐츠와 타겟국가에 따른 **예상 팔로워의 수를 예측**하여, 성장 가능성이 높은지를 판단하고자 합니다.

04 데이터 수집 및 전처리

데이터는 kaggle의 데이터셋을 이용

Data Analysis on Top Instagram Popular Influencers

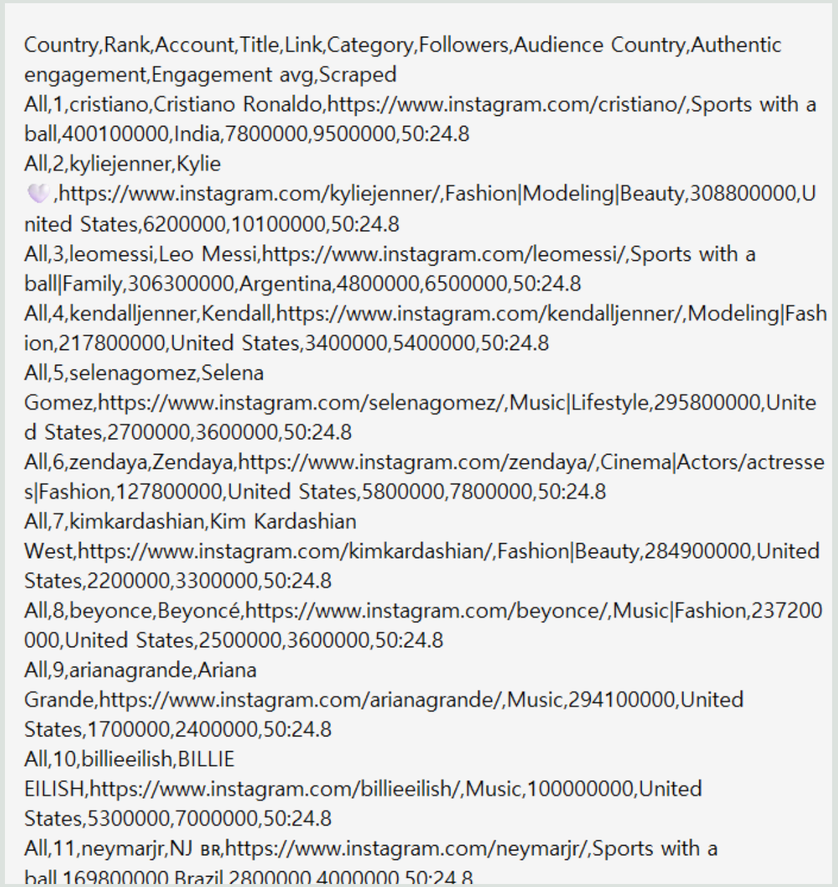
- https://www.kaggle.com/datasets/ankulsharma150/marketing-analytics-project
- 해당 데이터셋에서는 특정 컨텐츠의 주제가 많이 소비 될 것으로 예상되는 지역과 실제 데이터의 차이를 이야기함.
- 인스타그램 계정 중 팔로워 수가 가장 많은 컨텐츠는 구기종목. 하지만 축구와 관련이 깊은 영국에서의 팔로우 비중은 상위권에 속하지 못함
- 따라서 컨텐츠에 따른 **팔로워의 국가를 분석하여 타겟국가를 선정** 하고자 함.

실제 머신러닝 진행

01 데이터 전처리

원본 데이터셋에 포함되어 있던 항목은 다음과 같습니다.

- Country: 국가
- Rank: 순위
- Account: 인스타그램 계정
- Title: 계정(소유자)의 이름
- Link: 인스타그램 링크
- Category: 카테고리
- Followers: 팔로워 수
- Audience Country: 대상 국가
- Authentic engagement
  - 좋아요, 댓글, 공유와 같은 활동 중 인위적인 참여를 제외한 반응
- Engagement avg: 게시물당 평균적인 반응(좋아요, 댓글, 공유 등)
- Scraped: 데이터 수집 시점



Country,Rank,Account,Title,Link,Category,Followers,Audience Country,Authentic engagement,Engagement avg,Scraped
All,1,cristiano,Cristiano Ronaldo,https://www.instagram.com/cristiano/,Sports with a ball,400100000,India,7800000,9500000,50:24.8
All,2,kyliejenner,Kylie 🤍,https://www.instagram.com/kyliejenner/,Fashion|Modeling|Beauty,308800000,United States,6200000,10100000,50:24.8
All,3,leomessi,Leo Messi,https://www.instagram.com/leomessi/,Sports with a ball|Family,306300000,Argentina,4800000,6500000,50:24.8
All,4,kendalljenner,Kendall,https://www.instagram.com/kendalljenner/,Modeling|Fashion,217800000,United States,3400000,5400000,50:24.8
All,5,selenagomez,Selena Gomez,https://www.instagram.com/selenagomez/,Music|Lifestyle,295800000,United States,2700000,3600000,50:24.8
All,6,zendaya,Zendaya,https://www.instagram.com/zendaya/,Cinema|Actors/actresses|Fashion,127800000,United States,5800000,7800000,50:24.8
All,7,kimkardashian,Kim Kardashian West,https://www.instagram.com/kimkardashian/,Fashion|Beauty,284900000,United States,2200000,3300000,50:24.8
All,8,beyonce,Beyoncé,https://www.instagram.com/beyonce/,Music|Fashion,237200000,United States,2500000,3600000,50:24.8
All,9,arianagrande,Ariana Grande,https://www.instagram.com/arianagrande/,Music,294100000,United States,1700000,2400000,50:24.8
All,10,billieeilish,BILLIE EILISH,https://www.instagram.com/billieeilish/,Music,100000000,United States,5300000,7000000,50:24.8
All,11,neymarjr,NJ ʙʀ,https://www.instagram.com/neymarjr/,Sports with a ball,169800000,Brazil,2800000,4000000,50:24.8

01 데이터 전처리

그 중에서 필요한 항목만 남기고 글자를 숫자로 변환하는 전처리 과정을 진행하였습니다.

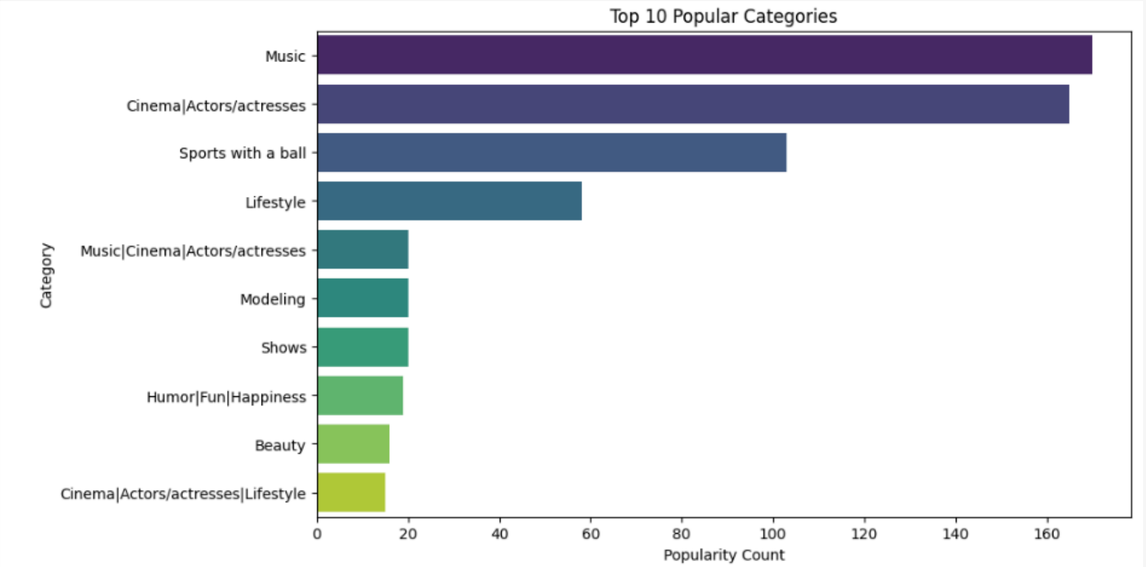
	Rank	Title	Category	Followers	Audience Country	Authentic engagement	Engagement avg
0	1	Cristiano Ronaldo	Sports with a ball	400100000	India	7800000	9500000
1	2	Kylie ♡	Fashion|Modeling|Beauty	308800000	United States	6200000	10100000
2	3	Leo Messi	Sports with a ball|Family	306300000	Argentina	4800000	6500000
3	4	Kendall	Modeling|Fashion	217800000	United States	3400000	5400000
4	5	Selena Gomez	Music|Lifestyle	295800000	United States	2700000	3600000

```python
import pandas as pd

# 데이터 로드
data = pd.read_csv("/kaggle/input/instainfluencer/Instagram.csv")

# 불필요한 열 제거 (필요에 따라)
# 'Followers' 열이 문자열인지 확인 후 전처리
if data['Followers'].dtype == 'O':  # dtype 'O'는 object 타입을 의미 (일반적으로 문자열)
    data['Followers'] = data['Followers'].str.replace(",", "").astype(int)
else:
    data['Followers'] = data['Followers'].astype(int)

# 다른 열들도 숫자로 변환
data['Authentic engagement'] = data['Authentic engagement'].astype(int)
data['Engagement avg'] = data['Engagement avg'].astype(int)

# 'Country', 'Account', 'Link', 'Scraped' 열 삭제
data = data.drop(columns=["Country", "Account", "Link", "Scraped"])

data.head()
```

02 데이터 시각화

상위 10개 카테고리

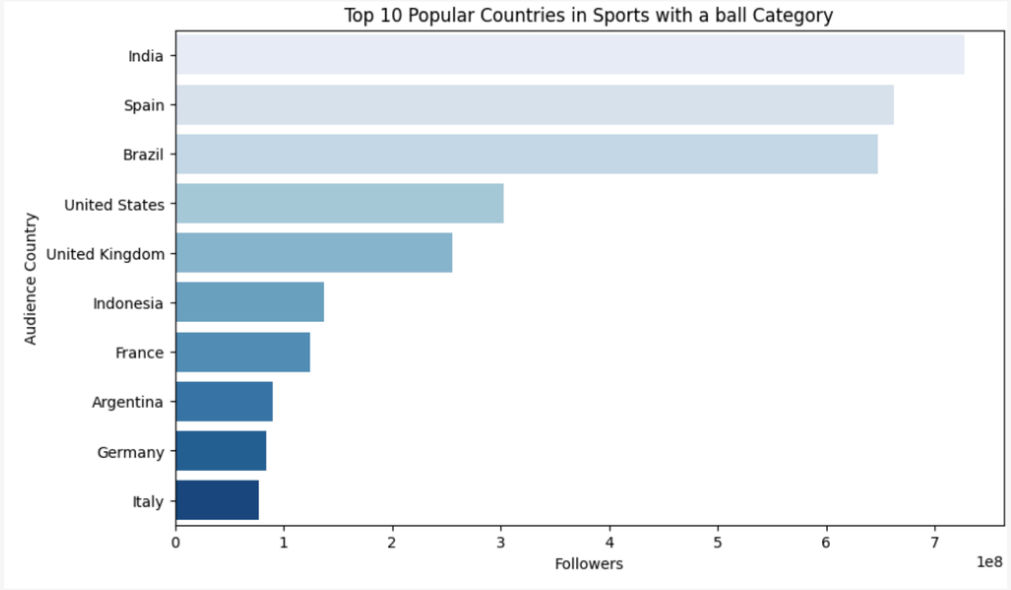
```python
import pandas as pd

# 데이터 로드
data = pd.read_csv("/kaggle/input/instainfluencer/Instagram.csv")  # 파일 경로를 지정해주세요

# 카테고리별로 전체 등장 횟수 계산
category_counts = data['Category'].value_counts().reset_index(name='Count')
category_counts.columns = ['Category', 'Count']

# 상위 10개 카테고리 필터링
top10_categories = category_counts.head(10)

# 시각화
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.barplot(x='Count', y='Category', data=top10_categories, palette='viridis')
plt.title('Top 10 Popular Categories')
plt.xlabel('Popularity Count')
plt.ylabel('Category')
plt.show()
```



02 데이터 시각화

가장 인기있는 카테고리인 축구의 청중국가 분포

```
# 카테고리별로 인기 있는 국가와 그에 해당하는 청중 수 계산
category_country = data.groupby(['Category', 'Audience Country'])['Followers'].sum().reset_index()

# 특정 카테고리 예시로 필터링 (예: 'Sports with a ball')
category_country_filtered = category_country[category_country['Category'] == 'Sports with a ball']

# 시각화
plt.figure(figsize=(10, 6))
sns.barplot(x='Followers', y='Audience Country', data=category_country_filtered.sort_values('Followe
plt.title('Top 10 Popular Countries in Sports with a ball Category')
plt.xlabel('Followers')
plt.ylabel('Audience Country')
plt.show()
```



03 모델 학습

Pycaret을 이용하여 모델의 결정 및 학습을 진행하였습니다.
다양한 모델을 평가한 결과, Light Gradient Boosting Machine (LightGBM) 모델이 선택되었고
R^2 값이 0.7929로, 예측 정확도가 79%였습니다.

```python
from pycaret.regression import *
import pandas as pd

# 전처리된 데이터 로드
data = pd.read_csv('/content/preprocessed_data.csv')

# PyCaret 환경 설정 (타겟 변수 'Followers'로 설정)
exp = setup(data, target='Followers', categorical_features=['Category', 'Audience Country'], session_id=123)

# 모델 비교 후 선택
best_model = compare_models()

# 모델 학습
final_model = create_model(best_model)

# 모델 튜닝
tuned_model = tune_model(final_model)

# 예측
predictions = predict_model(tuned_model, data)

# 예측된 값 확인
print(predictions.head())
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Light Gradient Boosting Machine	4612437.9725	310267911844320.5625	17614423.4037	0.7929	0.2749	0.1222

04 예측

Pycaret을 이용하여 모델의 결정 및 학습을 진행하였습니다.
Category, Audience Country을 통해 Followers을 예측하고자 했으며,
다양한 모델을 평가한 결과, Light Gradient Boosting Machine (LightGBM) 모델이 선택되었고
R^2 값이 0.2527로, 예측 정확도가 약 25%였습니다.

```python
from pycaret.regression import *
import pandas as pd

# 전처리된 데이터 로드
data = pd.read_csv('/content/preprocessed_data.csv')

# PyCaret 환경 설정 (타겟 변수 'Followers'로 설정)
exp = setup(data, target='Followers', categorical_features=['Category', 'Audience Country'], session_id=123)

# 모델 비교 후 선택
best_model = compare_models()

# 모델 학습
final_model = create_model(best_model)

# 모델 튜닝
tuned_model = tune_model(final_model)

# 예측
predictions = predict_model(tuned_model, data)

# 예측된 값 확인
print(predictions.head())
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Gradient Boosting Regressor	15886455.0921	1119316761043811.3750	33456191.6698	0.2527	0.7961	0.9925

04 예측

정확도가 낮아 Category, Audience Country, Followers을 통해 Engagement avg을 예측하는 모델을 학습
다양한 모델을 평가한 결과, Extra Trees Regressor모델이 선택되었지만, 이 모델 역시 정확도가 낮아 사용할 수 없을 것으로 보입니다.

```
from pycaret.regression import *
import pandas as pd

# 데이터 불러오기
data = pd.read_csv('/content/preprocessed_data.csv')

# Engagement avg 예측
exp = setup(data, target='Engagement avg', categorical_features=['Category', 'Audience Country'], session_id=123)

# 모델 비교 후 선택
best_model = compare_models()

# 선택된 모델 학습
final_model = create_model(best_model)

# 모델 튜닝
tuned_model = tune_model(final_model)

# 예측
predictions = predict_model(tuned_model, data)

# 예측된 값 확인
print(predictions.head())
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Extra Trees Regressor	529778.5177	1041876705030.3785	1020723.6183	0.3509	0.8386	1.2269

05 결과

학습한 두 모델 모두 정확도가 낮아 사용할 수 없었습니다.
학습에 사용된 데이터량이 매우 적고(1000개), 상위 인플루언서에 집중되어 있는 데이터이기 때문에
이러한 결과가 나타난 것 같습니다.
추후 더 적합한 데이터셋을 통하여 학습을 진행하도록 하겠습니다.

Thank you

감사합니다.