

데이터마이닝 기법을 적용한 인플루언서 분석 및 선정 방안 제시

요조

산업공학과 2015147001 최동훈

산업공학과 2015147031 이정현

산업공학과 2017147015 곽지운

산업공학과 2017147036 박현준

지도교수 손소영
조교 우현우

Contents



YONSEI UNIVERSITY

Introduction

Literature
Review

Methodology
& Data

Analysis

Conclusion

Abstract

인플루언서 마케팅이란 SNS와 온라인 커뮤니티에서 많은 팔로워 수를 가진 인플루언서를 통해 진행되는 마케팅을 의미한다. 인플루언서들은 팔로워에 따라 1~10K의 규모를 가진 Nano인플루언서, 10~50K의 Micro인플루언서, 50~500K의 Mid-tier인플루언서, 500K~1M의 Macro인플루언서, 1M+의 Mega인플루언서로 구분된다. 이들은 팔로워 총수를 통해 SNS상에서 막강한 영향력을 발휘하고 이러한 장점을 바탕으로 기업은 자사의 브랜드 인지도 형성, 효율적인 소비자 타겟팅에 활용하고 있는 추세이며 관련 시장은 지속해서 성장하고 있다.

현재 인플루언서 기반 마케팅에서 인플루언서의 팔로워 수, 댓글 수등의 판단 지표를 바탕으로 기업내 담당자들의 주관적인 판단으로 주요한 의사결정이 이루어지고 있다. 이러한 의사결정은 인스타그램 정보공개의 제한성, 맞춤화된 지표의 부족 등으로 인해 신뢰성이 떨어지는 문제가 있다. 또한 팔로워 수는 적지만 팔로워들의 반응도가 높은 나노인플루언서의 등장으로 여러 판단 지표의 종합적인 판단이 필요한 상황이다.

따라서 본 연구에서는 상기된 문제들을 해결하고 인플루언서의 영향력을 판단하는 정량적인 지표를 제시하기 위해 Factor Analysis를 통해 영향력을 가진 지표들의 구성을 도출하고 Clustering을 통해 비용효율적인 마케팅 집행 방안을 소개한다. 또한, CBR을 통해 기업이 새로운 인플루언서를 통해 마케팅을 진행 시 앞서 소개한 방안이 적용될 수 있는지에 대해 검증한 결과를 제시한다.

SNS 사용자의 증가로 채널을 활용하는 기업들의 소셜 마케팅 활용이 증가하고 있으며,
그 중심에는 전통적인 Celebrity 이외에 새로운 영향력 행사자인 인플루언서가 존재합니다



인플루언서 마케팅이란?

SNS 사용이 많은 소비자들에게 막강한 영향을 미치는
인플루언서들을 통해, 다양한 SNS 채널에서
그들의 독창적인 콘텐츠를 활용하여
브랜드 홍보 등의 SNS 마케팅 진행

인플루언서 마케팅 부상 원인

기업 입장

- ✓ 브랜딩
 - 인플루언서 이용하여 브랜드 커뮤니케이션 진행
→ **브랜드 인지도 형성에 기여**
 - 인플루언서의 독창적 콘텐츠 활용
- ✓ 구체적인 소비자 타겟팅
 - SNS 사용자 대상의 구체적 광고 집행 가능
→ **정확한 소비자 파악 가능**

사용자 입장

- ✓ 친근감 있는 모델
 - 일반 Celebrity보다 친밀감과 공감대가 형성되는 인플루언서 팔로우하여 신뢰를 구축함
- ✓ 간접적인 제품 경험 가능
 - 제품 구매 이전 SNS을 통한 체험 검색
→ **인플루언서의 '리뷰어' 역할 활용**

출처: 브랜드 커뮤니케이션 활성화를 위한
효과적인 인플루언서(Influencer) 마케팅 전략 개발 제안 (문지원, 김원경. 2020)

인플루언서의 계층을 나눌 만큼 팔로워 수라는 주요 기준이 존재하지만, 인플루언서의 특성으로 평가하기도 하는 추세이며 이에 대한 정량적인 판단 기준은 부재한 상황입니다

인플루언서 구분

인스타그램 계층

- ✓ 팔로워 수를 기반으로한 계층 존재
- ✓ 단순 팔로워 수가 아니라도 광고효과 높은 사례 등장

인플루언서 유형

Mega	1백만명 이상
Macro	50만명 ~ 1백만명
Mid-Tier	5만명 ~ 50만명
Micro	1만명 ~ 5만명
Nano	1천명 ~ 1만명

나노 인플루언서 중요성 대두

- ✓ 팔로워 수는 적지만 팔로워들의 관심도, 참여도, 반응률이 높음. 따라서 비용 대비 높은 성과 기대 가능
- ✓ 개인적으로 친밀감이 높은 팔로워와 소통이 가능한 잠재 고객을 보유하고 있어 영향력 있는 사람들의 팔로워들보다 참여율이 높음

출처:
팔로워 충성도 높은 ‘나노 인플루언서’를 주목하라
(한국무역협회)

인플루언서 마케팅 문제상황

문제상황 1) 정성적인 판단에 기초한 현 업계 상황

- ✓ 실제 담당자들이 정량적인 지표가 부재한 채, 담당자 개인의 정성적인 판단만을 통해 인플루언서를 평가하고 있음
- ✓ 정량적인 지표는 사이트에서 제공하는 표면적인 정보(프로필 상의 정보) 이외 접근하기 어려움

출처:
모 쇼핑앱 마케팅 담당자 인터뷰



문제상황 2) 인스타그램에 맞춤화된 지표 부재

- ✓ 기존의 Facebook에서 사용된 개념을 그대로 적용한 비공식적인 지표 존재 (Engagement Rate)
- ✓ 인스타그램 특수성 반영X

Instagram Engagement Rate Formulas

There is no 'official' Instagram engagement rate formula, so choose one of the below methods depending on what you are using it for.

Instagram Engagement Rates*	Method 1	Method 2	Method 3
*Please note that these are not official results.	Likes + Comments / Total Followers x 100	Likes + Comments / Post Impressions x 100	Engagements / Reach x 100
	Useful to compare different Instagram accounts	Useful to compare different social networks	Useful to improve performance on Instagram

*Engagement rate is a percentage. Do not multiply by 100 if it is already done in the above equations.

출처:
2017 Social Media Industry Benchmark Report
(RivalIQ, 04 Apr.17)

데이터 마이닝 기법을 활용하여 인스타그램 내에서 수집 가능한 데이터를 통해
인플루언서를 판단하는 정량적 지표를 제시하는 것이 프로젝트의 목적입니다

연구 내용

■ 연구 목적

- ✓ 팔로워 수만이 절대적인 기준이 될 수 있는지에 대한 검증
- ✓ 인플루언서를 평가하는 정량적 지표 도출
- ✓ 팔로워 수에 의한 계층이 다른 인플루언서 간의 유사도를 파악하여 비용 측면에서 효율적인 마케팅 방안을 제시함

■ 연구 대상

- ✓ '뷰티', '패션' 카테고리의 인스타그램 인플루언서 대상으로만 한정 진행
- ✓ 독자적인 브랜드 런칭을 하지 않고, 또한 연예인 기반이 아닌 인플루언서
- ✓ 각 계층 별 인플루언서 명수 다르게 하여 총 80명

Tier (# of Followers)	인플루언서 수 (+test data)
Mid-Tier A (200k~500k)	10 (+3)
Mid-Tier B (50k~200k)	20 (+3)
Micro (10k~50k)	22 (+3)
Nano (1k~10k)	28 (+3)

1st RQ

팔로워 수와 채널 내에 확인할 수 있는 피상적인 정보
이외 판단 지표가 어떤 것이 있을까?

- ✓ 일반 사용자가 확인할 수 있는 정보는 피상적인 것에 제한이 있기 때문에 데이터 마이닝 기법을 적용하고자 함

2nd RQ

새로운 특성으로 기반으로 분류했을 때, 팔로워 수로 구분된
서로 다른 계층의 인플루언서가 동일한 분류가 될 수 있을까?

- ✓ 팔로워 수 계층이 달라도 유사한 특성을 갖는 인플루언서를 제시하여 비용효율적인 마케팅 집행을 제시할 수 있음

3rd RQ

앞서 분류를 진행한 후에, 해당 판단 기준과 분류가 기업이
원하는 새로운 인플루언서에도 적용할 수 있는가?

- ✓ 특정 인플루언서를 원할 경우 팔로워 수에 의한 노출 효과
이외에 해당 인플루언서의 특성을 제시할 수 있음



브랜드의 인플루언서 마케팅을 활용하는 논문을 통해 시장의 구조를 이해하였고,
신규 채널의 영향력을 행사하는 콘텐츠 제작자를 대상으로 하는 정량적인 판단 연구를 참고했습니다

선행 연구 1)

* 논문 명.

브랜드 커뮤니케이션 활성화를 위한 효과적인 인플루언서 마케팅 전략 개발 제안, 문지원, 김원경(2019)

* 논문 주요 내용

- ✓ 뉴미디어 시대에 기업과 소비자의 연결고리로서 인플루언서가 가지는 영향력을 국내외 기업 사례를 통해 분석함
- ✓ 소셜 미디어 내에서 인플루언서 출현 계기와 소비자와의 관계, 현재 인플루언서 마케팅 진행 단계를 분석함
- ✓ 기업의 브랜드 특성에 알맞은 인플루언서 선정, 검증 프로세스를 제안함

→ 소비자와의 원활한 커뮤니케이션을 위한 인플루언서 마케팅의 효과적인 전략 제안함

1) 인플루언서 분석

- ✓ 유명 셀럽보다 친근하고 편안한 콘텐츠
- ✓ 소비자와 팬덤 관계를 맺고 신뢰 형성

2) 인플루언서 마케팅 시장 분석

- ✓ 소비자와 활발한 커뮤니케이션을 통한 브랜드 가치 실현
 - ✓ 자사의 브랜드 이미지 전달 도구로 인플루언서 활용
- 인플루언서 마케팅이 기업의 마케팅 전략으로 떠오르고 있지만, 인플루언서 선정 기준이 미비하다는 점에 시사점이 있다.

선행 연구 2)

* 논문 명.

데이터 마이닝을 이용한 유튜브 인기 동영상 콘텐츠 분석, 김희숙(2020)

* 논문 주요 내용

- ✓ 유튜브 콘텐츠 제작자의 수익성 예측을 위해 정성적 데이터와 정량적 데이터 수집 및 분석
- ✓ 키워드의 정성적 분석을 통해 키워드와 주요 뉴스와의 관련성 도출함
- ✓ 조회수, 댓글 수, 좋아요 수, 싫어요 수에 대한 **정량적 분석을 통해 각 수치 간 상관관계 파악함**
- ✓ 인과관계 분석을 위해 회귀분석을 진행해 **수익(조회수)에 댓글 수, 좋아요 수, 싫어요 수가 미치는 영향력 파악함**

1) 정성적 분석

-영상과 주요 뉴스와의 관련성 → 수익과 양의 상관관계

2) 정량적 분석

- 좋아요 수, 싫어요 수, 댓글 수 → 수익과 양의 상관관계
 → 콘텐츠 제작자들의 영향력을 정량적 데이터를 통해 측정하고자 했다는 점에 시사점이 존재



전통적인 Celebrity를 판단하기 위한 평가기준을 설정하는 논문과,
해당 논문을 활용하여 인플루언서를 분석한 선행연구들을 분석하여 변수 설정에 참고했습니다

선행 연구 3)

* 논문 명.

Construction and Validation of a scale to Measure Celebrity Endorsers' Perceived Expertise, Trustworthiness, and Attractiveness
Roobina Ohanian(1990)

* 논문 주요 내용

- ✓ 셀러브리티 평가 지표를 설문조사로 진행함.
 - ✓ 설문조사 결과를 통해 평가 지표 15개를 Factor Analysis를 통해 특성을 도출함
 - ✓ 총 3가지 특성이 도출됨
(FA와 선행연구 분석을 통해)
- 평가 기준은 총 3가지 특성으로 정리됨

특성 1) Expertise = 전문성

- ✓ expert, experienced, knowledgeable, qualified, skilled

특성 2) Trustworthiness = 신뢰성

- ✓ dependable, honest, reliable, sincere, trustworthy

특성 3) Attractiveness = 매력성

- ✓ attractive, classy, beautiful, elegant, sexy

선행 연구 4)

* 논문 명.

인스타그램 인플루언서 마케팅의 팔로워 지각 효과에 관한 연구, 남연주, 김용호(2021)

* 논문 주요 내용

- ✓ 인스타그램 인플루언서 코어 팔로워 행태와 규모의 지각이 사용자의 소비자의 해당 인플루언서를 향한 구매 의도와 태도에 미치는 영향력을 판단함.
- ✓ 설문조사를 통해 인스타그램 인플루언서 구매의도 측정.
인플루언서를 판단하는 과정에서 논문1)의 3가지 주요 특성을 활용함
- ✓ 매력성의 경우 외모적 매력과 사회적 매력(소통여부 등)을 구분하여 사용

1) Attractiveness

- 외모적 매력
 - 양의 상관관계 없음
 - 사회적 매력
 - 양의 상관관계 존재
- 매력성을 외모적 매력과 사회적 매력을 구분하는데에 시사점이 존재

2) 전문성

- 양의 상관관계 존재

3) 진실성

- 양의 상관관계 존재



앞선 두 가지 선행연구를 참고하여 매력성, 전문성, 진실성 세가지를 주요 판단 기준 특성으로 참고하여 인스타그램 채널에서 확보할 수 있는 정보를 특성에 근거하여 도출하여 데이터의 독립변수로 지정했습니다

변수 도출 과정	Data Feature 구성
<p>전통적 Celebrity 평가 기준</p> <ul style="list-style-type: none"> ✓ Attractiveness (사회적 매력) ✓ Expertise ✓ Trustworthiness  <p>인스타그램 SNS 특성 반영</p> <ul style="list-style-type: none"> ✓ Attractiveness → 사회적 매력 → Follower 기반 특성 <div style="border: 1px dashed gray; padding: 10px; margin-top: 10px;"> <p>소비자와의 소통과 셀럽 팔로워의 커뮤니티에서 발생하는 소통 여부가 사회적 매력에 해당되는 요소</p> </div> <ul style="list-style-type: none"> ✓ Expertise → 게시물 + 광고분야 기반 특성 ✓ Trustworthiness → 인플루언서 개인의 특성 <p>(출처) Construction and Validation of a scale to Measure Celebrity Endorsers' Perceived Expertise, Trustworthiness, and Attractiveness Roobina Ohanian (1990) 인스타그램 인플루언서 마케팅의 팔로워 지각 효과에 관한 연구 남연주, 김용호(2021)</p>	<p>1) 기본 특성 (인구통계학+표면정보)</p> <ul style="list-style-type: none"> ✓ 성별 ✓ 나이대 ✓ # of followers ✓ # of 게시글 ✓ Avg. Comment # ✓ Avg. Likes # ✓ Youtube 활용 여부 ✓ 타 플랫폼 활용 여부 <p>2) Expertise</p> <ul style="list-style-type: none"> ✓ 광고게시물#/2021년 게시물 ✓ 최근 1개월 간 게시물 개수 ✓ 최근 1개월 간 광고 게시물 개수 ✓ 광고 게시물의 평균 길이 ✓ 광고 게시물의 평균 Hashtag 수 ✓ 동일 브랜드 광고 반복 여부 ✓ 광고 게시글의 # of Likes / # of Followers ✓ 광고 게시글의 # of Comments / # of Followers ✓ 비광고 게시글의 # of Likes / # of Followers ✓ 비광고 게시글의 # of Comments / # of Followers <p>3) Trustworthiness</p> <ul style="list-style-type: none"> ✓ Negative 논란 발생 여부 <p>4) Attractiveness</p> <ul style="list-style-type: none"> ✓ 2020년 5월부터 현재까지 (인플루언서 가 대댓글 단 게시물 수) / (게시물 수) ✓ 2020년 5월부터 현재까지 (인플루언서 가 단 평균 대댓글 개수) / (기간 내 평균 댓글수) ✓ 최근 3개월 동안 (반복해서 댓글 단 사람 수)/(게시물 평균 댓글 단 사람 수)
	<p>웹 크롤링을 활용하여 인스타그램 내의 데이터 수집</p>

각 계층 별로 총 80명의 Train set 인플루언서와 12명의 Test set 인플루언서를 선정하였고,
Factor Analysis, Clustering, CBR(KNN모델)를 통하여 각 RQ를 해결하며 연구를 진행했습니다

연구 대상					
	Nano	Micro	Mid-Tier B	Mid-Tier A	
Train	boilylook_j_botong ... _sic_h	eunkyung302 binnybaby_ ... milkcat90	yeaseullee Jenvly ... _seoyeonn	o62oo cheristyle_ ... yoo.xx	
Data #	28명	22 명	20 명	10 명	
Test	_byeoolee Thenanamilk gyo_owo_c	hyon_ing lovehean_ d.o.j.e.e	chaevely__ su_xy _s_yj_	woo_gi_ so0o_h_ ghyun_	
Data #	3 명	3 명	3 명	3 명	

→ Train Data 80 명, Test Data 12 명 데이터 추출

1st RQ

팔로워 수와 채널 내에 확인할 수 있는 피상적인 정보
이외 판단 지표가 어떤 것이 있을까?

- ✓ Factor Analysis를 통해 Feature가 어떻게 묶이는
지 확인

2nd RQ

새로운 특성을 기반으로 분류했을 때, 팔로워 수로 구분하는
서로 다른 계층의 인플루언서가 동일한 분류가 될 수 있을까?

- ✓ 계층을 새로운 특성으로 구분하여 Clustering 기법
을 통한 새로운 군집 확인

3rd RQ

앞서 분류를 진행한 후에, 해당 판단 기준과 분류가 기업이
원하는 새로운 인플루언서에도 적용할 수 있는가?

- ✓ Clustering이 어떻게 나뉘는지 CBR을 통해 Test
데이터에 대한 예측



FA를 진행하기 앞서 변수를 선별하고 Factor의 수를 결정한 후
4가지 Rotation 방식을 비교해 데이터에 가장 적합한 방식으로 FA를 진행했습니다

Feature Drop

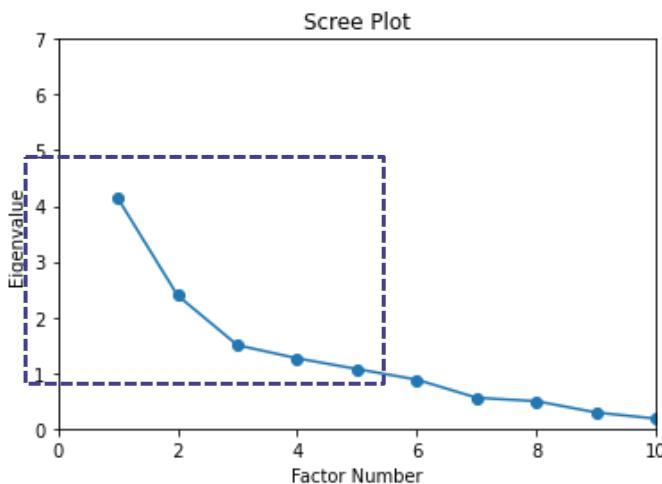
1) 기본 정보

- ✓ 성별, 나이, 팔로워 수, 포스팅 개수, 평균 좋아요
- ✓ 본 과제는 기본 정보를 제외한 판단 요소를 찾고자 함에 목적이 있기에 FA 과정에서 제외함

2) 논란 여부, 타 플랫폼 및 유튜브 활용 여부

- ✓ 80명 중 논란 있는 인플루언서 2명
- ✓ 결과에 영향을 줄만큼 양이 많지 않아 제외

Factor 수 결정



- ✓ Eigen value값이 1 이상 기준으로 Factor 개수 결정
→ Factor 개수 5개

Rotation 방식 비교

1) Varimax Rotation

- ✓ 직교 회전
- ✓ 행렬의 열 기준 분산 극대화
- ✓ Cumulative variance: 0.706848
- 전체 분산의 약 70% 정도의 설명력을 가진다

2) Quartimax Rotation

- ✓ 직교 회전
- ✓ 행렬의 행 기준 분산 극대화
- ✓ Cumulative variance: 0.706848
- 전체 분산의 약 70% 정도의 설명력을 가진다

3) Promax Rotation

- ✓ Varimax 기반 사각 회전
- ✓ 요인 사이 낮은 상관성을 갖도록 함
- ✓ Cumulative variance: 0.711016
- 전체 분산의 약 71% 정도의 설명력을 가진다

4) Oblimin Rotation

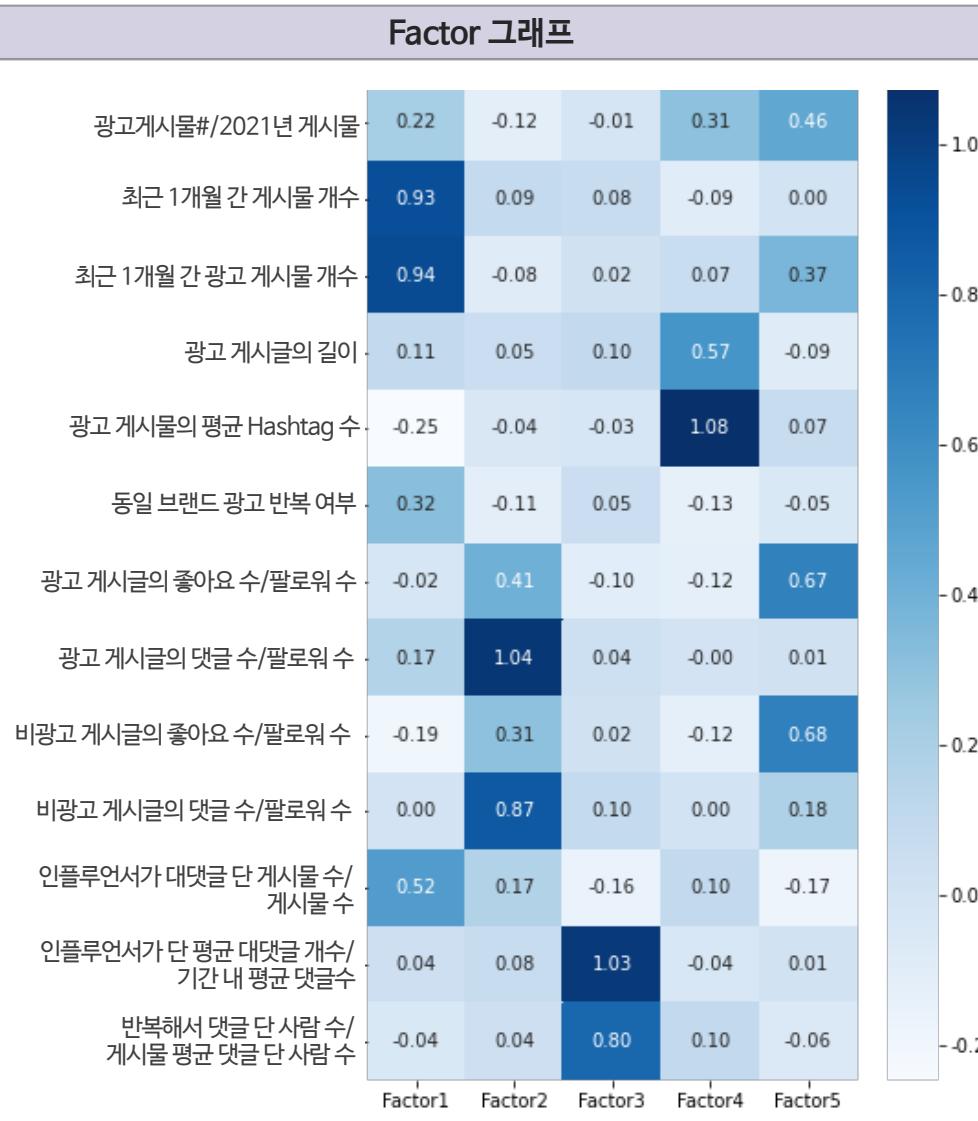
- ✓ 사각회전
- ✓ 요인의 상관성을 인정하고, 없을 경우 직교회전으로 진행됨
- ✓ Cumulative variance: 0.665140
- 전체 분산의 약 66% 정도의 설명력을 가진다

→ Cumulative variance가 가장 큰 Promax Rotation으로 결정



Promax Rotation 기준 Factor 그래프를 도출하였고,
각 요인 별로 유의미한 변수를 선정하여 각 요인의 의미를 해석했습니다

Factor 그래프



Factor 분석

1) Factor 1

- ✓ 최근 1개월 간 게시물 개수, 최근 1개월 간 광고 게시물 개수가 유의미한 요인
→ **최근 1개월간 게시글 업로드 정도를 의미하는 Factor**

2) Factor 2

- ✓ 광고 게시물의 팔로워수 대비 댓글 수, 비광고 게시물의 팔로워수 대비 댓글 수가 유의미한 요인
→ **댓글 Engage 정도를 의미하는 Factor**

3) Factor 3

- ✓ 인플루언서가 대댓글 단 게시물 수/게시물 수, 반복해서 댓글 단 사람 수/게시물 평균 댓글 단 사람 수가 유의미한 요인
→ **인플루언서 대댓글 정도를 의미하는 Factor**

4) Factor 4

- ✓ 광고 게시물의 길이와 광고 게시물의 평균 태그 개수가 유의미한 요인
→ **광고 게시물의 정교함을 의미하는 Factor**

5) Factor 5

- ✓ 광고 게시물의 팔로워수 대비 좋아요 수, 비광고 게시물의 팔로워수 대비 좋아요 수가 유의미한 요인
→ **좋아요 Engage 정도를 의미하는 Factor**



5가지 요인을 해석한 결과 피상적인 정보 이외 판단 지표임을 확인하여 RQ#1을 해결할 수 있었고, 선행연구와 비교 결과 유의미한 지표임을 확인할 수 있었습니다

FA를 통한 1st RQ 해결 과정

1st RQ

팔로워 수와 채널 내에 확인할 수 있는 피상적인 정보 이외
판단 지표가 어떤 것이 있을까?

- ✓ Factor Analysis를 통해 Feature가 어떻게 묶이는지 확인

✓ 5가지 팔로워 수 이외에 유의미한 Factor 도출

- 최근 1개월간 게시글 업로드 정도를 의미하는 Factor
- 댓글 Engage 정도를 의미하는 Factor
- 인플루언서 대댓글 정도를 의미하는 Factor
- 광고 게시물의 정교함을 의미하는 Factor
- 좋아요 Engage 정도를 의미하는 Factor

- ✓ Factor Analysis를 통해 팔로워 수와 같은 피상적인 정보가
아닌 새로운 인플루언서 판단 특성 5가지 도출



첫번째 RQ 해결

선행 연구와 Factor간 연관성

선행 연구 특성과의 공통점

Expertise (전문성)	Attractiveness(사회적매력)
업로드 활발함 (Factor 1)	대댓글 여부, 비율 (Factor 3)
광고글의 정교함 (Factor 4)	

- ▶ ✓ 분석 시 제외한 진실성 이외 전문성, 사회적 매력성
두 가지 특성에 부합한 요인 도출된 것으로 확인됨

선행 연구 특성과의 차이점

✓ 팔로워의 Engage 정도 관련 요인 도출

Factor 2	댓글 Engage 정도
Factor 5	좋아요 Engage 정도

- ✓ 팬덤의 Engage 정도를 측정 가능한
소셜 마케팅 특성반영

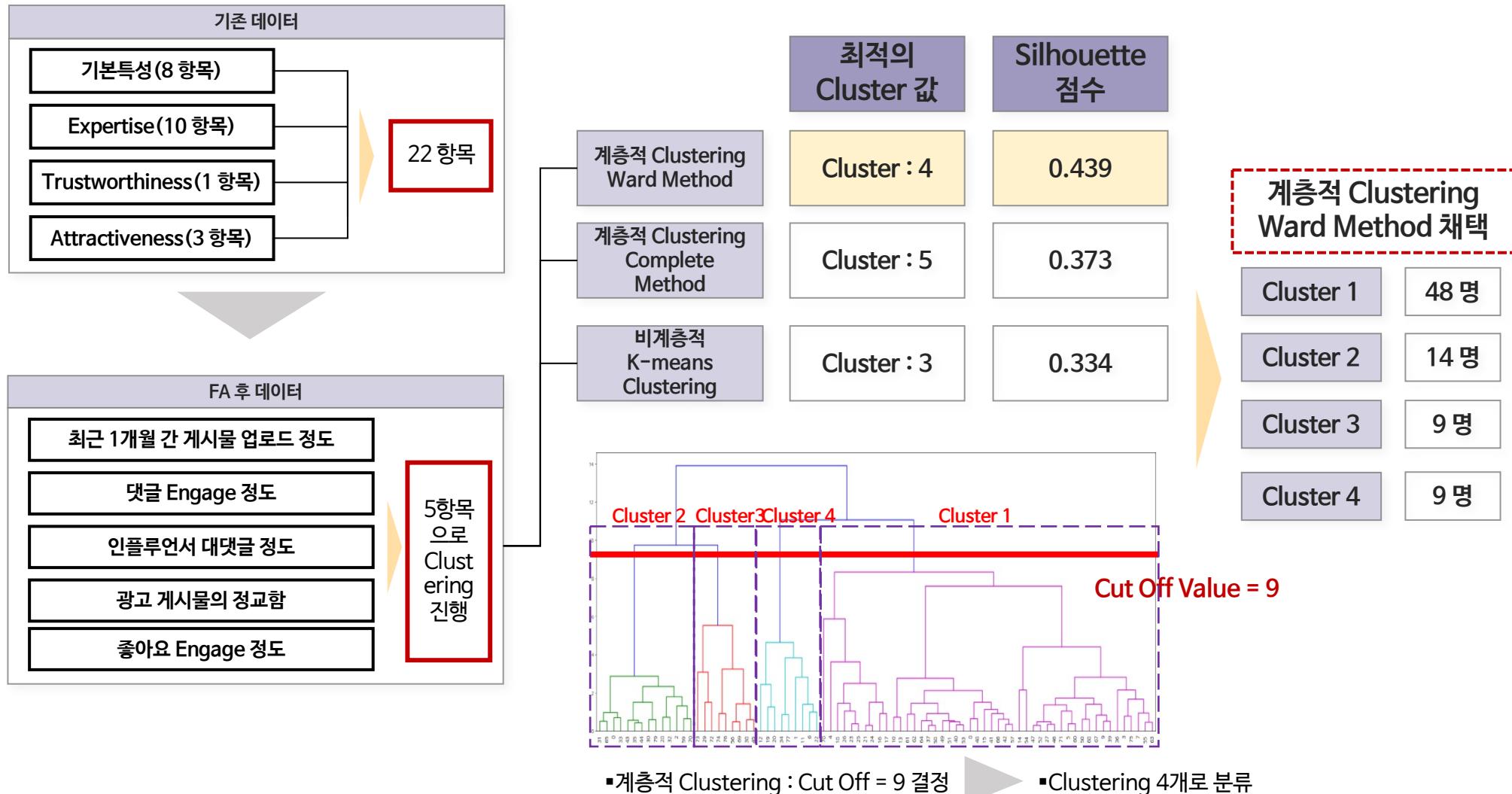
▶ 셀럽 관련 선행연구와
차이점 존재

- ✓ 다른 정도의 Engage Acting인
좋아요와 댓글이 분리된 개념
- ✓ 최근 게시물로 한정된 수치

} 기존 지표와
차이점 존재



앞선 5개의 요인에 대해 계층적 Clustering(Ward, Complete), 비계층적 Clustering을 진행하였고,
그 결과 Ward 방법을 이용한 계층적 Clustering 방법이 가장 실루엣 스코어가 높게 나왔습니다



가장 좋은 실루엣 스코어를 기록한 Ward 방식의 계층적 Clustering 결과로 4개의 Cluster에 대해 다음과 같이 해석했습니다

Cluster #	Data 수
1	14
2	9
3	9
4	48

recent 1 month post upload degree 댓글 Engage 정도 인플루언서 대댓글 정도 광고 게시물의 정교함 좋아요 Engage 정도

hc_cluster

1	-0.445702	-0.018507	1.485593	-0.350713	-0.521692
2	-0.931552	2.232416	1.032606	1.260666	0.236264
3	1.019635	-1.093581	-0.591362	-0.399059	2.207681
4	0.113481	-0.208134	-0.516031	-0.059260	-0.306080

Cluster 해석

	최근 1개월간 게시물 업로드 정도	댓글 Engage 정도	인플루언서 대댓글 정도	광고 게시물 의 정교함	좋아요 Engage 정도
Cluster 1	-	-	매우 높음	-	낮음
Cluster 2	매우 낮음	매우 높음	매우 높음	매우 높음	-
Cluster 3	매우 높음	매우 낮음	낮음	-	매우 높음
Cluster 4	-	-	낮음	-	-

■ 이때, 최근 1개월간 게시물 업로드 정도는 manual로 확인한 결과
모든 인플루언서들이 1~2일에 하나 정도는 올리는 상황에서
수치가 높은 인플루언서들은 하루에 n개씩 올리는 것이라서 **상대적인 개념임**

Cluster 1

- ✓ 사회적 매력이 강한(소통이 높음) 인플루언서들이지만,
팔로워들의 좋아요 engage는 상대적으로 낮은 편이다

Cluster 2

- ✓ 팔로워들의 댓글 engage 정도가 가장 높고 동시에 소통도
강하고 광고게시물을 올릴 때 정교하게 올려 전문성도 높음.
하지만 상대적으로 최근 업로드 정도가 낮다.

Cluster 3

- ✓ 좋아요 engage가 매우 높은 인플루언서. 대댓글 정도는
낮지만 최근 게시물 업로드는 활발한 편

Cluster 4

- ✓ 제일 많은 군집이며 계층이 다양하고, 절대값이 상대적으로
모두 다 낮은 편이고 차이를 많이 볼 수는 없다. 다른 군
집에 비해 소통이 좀 적은 편이다.



군집화 결과 서로 다른 계층의 인플루언서들이 동일 분류로 되어있어 두번째 RQ를 해결할 수 있었고,
군집을 바탕으로 구체적인 필요에 따라 마케팅 방안이 제시될 수 있었습니다

군집화를 통한 2nd RQ 해결 과정

2nd RQ

새로운 특성을 기반으로 분류했을 때, 팔로워 수로 구분하는
서로 다른 계층의 인플루언서가 동일한 분류가 될 수 있을까?

	Nano	Micro	Mid-B	Mid-A
군집1	2	6	3	3
군집2	0	2	6	1
군집3	8	1	0	0
군집4	18	13	11	6

- ✓ 군집의 특성에 따라 계층의 인플루언서가 섞여있음을 확인 할 수 있음
- ✓ 군집3의 경우 Engage가 높은데 이는 팔로워 수가 낮아도 참여율이 높은 인플루언서임을 확인함

- ✓ 서로 다른 계층의 인플루언서들이 팔로워 수가 아닌 다른 특성으로 구분된 군집에 동일한 분류화가 됨을 확인함



두번째 RQ 해결

새로운 군집화의 예시

마케팅 방안 제시 예시

1) 브랜딩이 필요하여 소통이 중요한 브랜드

- ✓ 팔로워들의 소통이 강한 군집1에 해당되는 인플루언서 대상으로 집행할 수 있음
- ✓ 이때, 충성도 높은 타겟 고객이 필요할 경우 군집2의 인플루언서 대상으로도 집행 가능함

2) 당장의 노출효과가 중요한 브랜드

- ✓ 팔로워 수는 적지만 최근 게시물 업로드가 가장 활발하고, 좋아요 Engage가 높아 노출 가능성이 높은 군집 3에 해당되는 인플루언서 대상 집행 가능
- ✓ 기존 팔로워 수 기반 계층을 이용하는 것도 방법

3) 비용 효율 확보 가능

- ✓ 군집4에 해당되는 인플루언서들 대상 집행할 경우 Mid-TierA 3명 선정을 **Mid-Tier A 2명과 Nano, Micro 2~3명**, 총 5명으로 구성하여 비용 효율 확보 가능함



예측 모델을 적용하기 위하여 앞선 군집을 범주형 종속변수로 지정하여 Train set을 재설정한 후,
전처리 과정을 진행하여 KNN모델 적용을 위한 최적 방식을 분석했습니다

Train Data 설정 및 전처리

FA & Clustering을 통해 Training Data의 Categorical Target Variable을 설정

최근 1개월 댓글 Engаж 인플루언서 광고 계시 좋아요 Encluster
0 -0.42664 0.004884 0.632604 -0.19241 -0.71443
1 0.400822 -1.23984 1.126466 0.96174 2.318124
2 0.065949 -0.21262 0.789285 -0.47011 -0.91874
3 0.188471 -0.31654 0.636072 0.69185 0.595476

- ✓ 5Factor Analysis & Clustering을
통한 기본 데이터 생성

- ✓ 기존 22 Features Dataset에 앞의 과정에서 나온 Cluster를 Categorical Target Variable로 설정

Column Drop, Scaling, Shuffling, Oversampling → KNN모델학습 전 전처리

	adposting	number_p	number_at_ad	postingad	동일 광고	광고 게시사	광고 게시자	비광고 게	비광고 게	Comment	Re_cmt_pk	Re_cmt_nu	cluster
0	0.623377	21	13	325.9231	6.85	1	0.011986	0.001012	0.012327	8.06E-05	4.512262	0.414286	0.077128
1	0.411765	4	6	235	15	1	0.133584	0.000859	0.220483	0.001828	1.857143	0.571429	0.069444
2	0.075	12	0	475	0	0	0.024199	0.000669	0.020144	0.000443	2.91	0.457143	0.100865
3	0.423077	16	9	275.8889	8.222222	1	0.079699	0.001239	0.09572	0.001216	6.847251	0.4	0.033453

- ✓ 22개 Features Dataset에서 ['sex', 'age', …] 등 총 9개 Feature를 Drop하여 13개 Feature를 가진 Dataset 설정

• (Feature, Target 분리, Scaling 및 Shuffling)

```
y.value_counts() # oversampling, SMOTE from imblearn.over_sampling import SMOTE X_resampled, y_resampled = SMOTE(random_state=2021).fit_resample(X_shuffled, y_shuffled) y_resampled.value_counts()
```

- ✓ 범주형 변수의 Imbalanced Data 문제를 해결하기 위해 Oversampling의 일종인 SMOTE 적용

→ Category별 44개의 Data 도출

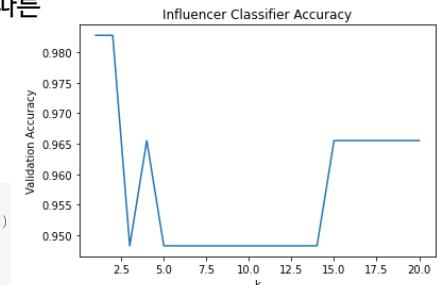
KNN모델 K-neighbors 설정 과정

Euclidean Metric

- ✓ Training Data, Validation Data 7:3 분리
 - ✓ Data간 거리 측정시 피타고라스 정리에 따른
최단거리 계산을 적용하는 유clidean 기준

$$\text{Euclidian distance} = \sqrt{\sum_{j=1}^J (x_j - y_j)^2}$$

```
for i in range(1,20):  
    knn = KNeighborsClassifier(i, weights="distance", metric="euclidean")  
    knn.fit(training_data, training_labels)  
  
    train_scores.append(knn.score(training_data, training_labels))  
    test_scores.append(knn.score(validation_data, validation_labels))
```

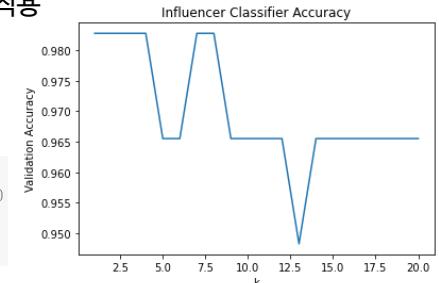


Manhattan Metric

- ✓ Training Data, Validation Data 7:3 분리
 - ✓ Data간 거리 측정시 블록방식의 계산을 적용하는 맨하탄 기준

$$\text{Manhattan distance} = \sum_{j=1}^J |x_j - y_j|$$

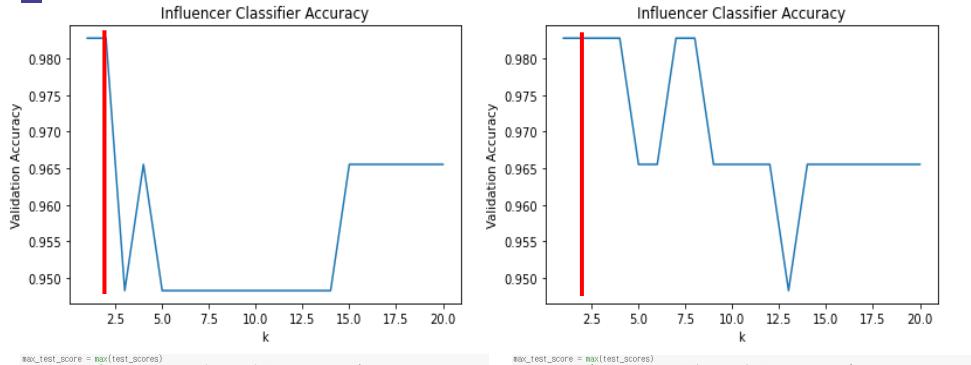
```
for i in range(1,20):  
    knn = KNeighborsClassifier(i, weights="distance", metric="manhattan")  
    knn.fit(training_data, training_labels)  
  
    train_scores.append(knn.score(training_data, training_labels))  
    test_scores.append(knn.score(validation_data, validation_labels))
```



각 Metric의 Accuracy를 비교하여 최적의 KNN모델을 결정한 후 Test Data에 대해 모델링을 진행했고, 최적 모델을 통해 Test Data의 결과를 예측하고 해석 및 활용방안 예시를 제시하였습니다

Test Data 예측 과정

Euclidean Metric & Manhattan Metric



Max test score 98.2758620689551 % and k = [1, 2]

Max test score 98.2758620689551 % and k = [1, 2, 3, 4, 7, 8]

→ Euclidean Metric에서 K=2인 경우가 가장 높은 Accuracy를
보이기에 KNN모델 기준으로 채택

▼ (Test Data 12개 전처리 및 KNN모델 학습)

```
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 2, weights="distance", metric="euclidean")
array([1, 4, 1, 4, 4, 4, 2, 2, 1, 3, 4, 1])
```

sex	age	number of number of 댓글수	like_averge/youtube	티 푸팅률	adposting.number_p_number_ad_posting	동일 콩고	광고 게시물	광고 게시물과 비광고 게시물	비광고 게시물	여부	Comment_Rate	Re_cmt_p	Re_cmt_r	cluster			
woo_gi_남자	20대	290784	415 95.54545	12342	0	1 0.267717	50	21 196.2857	9.857143	0 0.036842	4.05-05	0.051132	3.57E-05	0 8.149254	0.583333	0.048611	1
sooo_h_여자	20대	291489	248 77.44615	8753	0	0 0.289474	18	5 222.8	10.5	0 0.02309	4.09E-06	0.030337	3.97E-05	0 5.677966	0.666667	0.055556	4
_ghyun_여자	20대	273815	680 36.1	3604	0	1 0.244681	46	18 159.5	8.5	0 0.007833	3.61E-05	0.013770	3.37E-05	0 8.258706	0.816667	0.081169	1
chaevely_여자	20대	131851	59 51.7	5860	0	0 0.05	26	1 558	12	0 0.040154	9.10E-06	0.036472	8.51E-05	0 4.760479	0.033333	0.02967	4
su_xy_여자	20대	160170	102 19.27277	4490	0	1 0.04	27	10 93.6	11.5	0 0.003175	6.24E-06	0.003024	1.59E-06	0 5.52139	0.293190	0.026275	4
_x_y_여자	20대	163227	595 4.363636	2354	0	0 0.1879	24	8 227.875	18.5	1 0.0092	2.21E-06	0.013342	2.76E-05	0 7.247664	0.1 0.020619	0 4	4
d.o.j.e.e_여자	20대	384 10.0	20 102.0	0	0 0.4447	88	37 91.18919	5.3	1 0.000292	1.02E-05	0.000292	0 0.000292	0 0.000333	0.000333	0 0.000716	0 1	
lovehean_여자	20대	42297	176 33.0	1490	0	0 0.59	40	29 138.5862	5.5	1 0.05417	0.00042	0.001027	0.001017	0 7.185007	0.816667	0.07716	2
hyper_in_여자	20대	23620	131 11.33	1246	0	0 0.192308	30	11 117.5455	2.7	1 0.001461	0.00019	0.002999	0.002932	0 5.308683	0.9 0.133663	1	1
_byseole_여자	20대	5349	168 20	876.41	0	0 0.02881	23	1 51	1	0 0.189604	0.00109	0.21605	0.001808	0 4.6375	0.581667	0.047112	3
theranom_여자	30대	5404	258 4 164.9737	0	0 0.289474	21	8 91.625	4.5	1 0.04021	0.000647	0.023531	0.000574	0 5.588235	0.033333	0.00995	4	
gyo_ovo_여자	10대	7950	35 16.83	95.38298	0	0 0.244681	46	18 159.5	1	0 0.117349	0.001221	0.005419	5.54E-05	0 3.925532	0.9 0.084906	1	

예측 해석

예측결과 해석

- Training Dataset Clustering 결과 및 Cluster별 Factor 특성 & CBR 예측 결과

Training	Nano	Micro	Mid-Tier B	Mid-Tier A	Test	Nano	Micro	Mid-Tier B	Mid-Tier A
Cluster 1	2	6	3	3	Cluster 1	1	1	0	2
Cluster 2	0	2	6	1	Cluster 2	0	2	0	0
Cluster 3	8	1	0	0	Cluster 3	1	0	0	0
Cluster 4	18	13	11	6	Cluster 4	1	0	3	1

hc_cluster	최근 1개월간 게시물을 업로드 정도	댓글 Engage 정도	인플루언서 대댓글 정도	광고 게시물의 정교함	좋아요 Engage 정도
1	-0.445702	-0.018507	1.485593	-0.350713	-0.521692
2	-0.931552	2.232416	1.032606	1.260666	0.236264
3	1.019635	-1.093581	-0.591362	-0.399059	2.207681
4	0.113481	-0.208134	-0.516031	-0.059260	-0.306080

➤ Ex1) Test Data의 sooo_h_

→ Mid-Tier A에 속하는 인플루언서이지만 댓글, 좋아요의 Engage정도가 적은 cluster4에 속하는 것으로 예측되어 광고의뢰시 신중을 기해야 할 것으로 보인다.

➤ Ex2) Test Data의 d.o.j.e.e & lovehean_

→ 비교적 팔로워수가 적은 Micro 인플루언서들이지만 댓글 Engage정도가 높고 대댓글 정도 및 광고게시물의 정교함이 높은 cluster2에 속하는 것으로 예측되어 광고의뢰에 적절하다고 판단된다.

인스타그램의 인플루언서에 대한 데이터를 바탕으로 3가지 데이터마이닝 기법을 적용하여 본 팀이 설정한 3가지 RQ를 해결하고 그에 따른 시사점 3가지를 제시했습니다

1st RQ

팔로워 수와 채널 내에 확인할 수 있는 피상적인 정보
이외 판단 지표가 어떤 것이 있을까?

2nd RQ

새로운 특성을 기반으로 분류했을 때, 팔로워 수로 구분
하는 서로 다른 계층의 인플루언서가 동일한 분류가 될
수 있을까?

3rd RQ

앞서 분류를 진행한 후에, 해당 판단 기준과 분류가 기업
이 원하는 새로운 인플루언서에도 적용할 수 있는가?

문제상황

문제상황 1) 정성적인 판단에 기초한 현 업계 상황

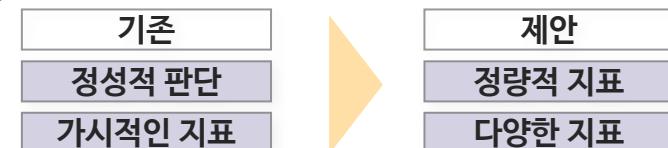
문제상황 2) 인스타그램에 맞춤화된 지표 부재

연구를 통한 해결

인플루언서 마케팅 상황에 대한 새로운 관점 제시

■ 시사점 1) 정량적이고 다양한 관점에서 판단할 수 있는
지표 제시

- ✓ 기존 논문에 기반한 인플루언서의 속성을 분석하여 수치
적으로 나타난 지표에 대한 제공
- ✓ 최근 게시물의 활동성, 소통여부 등에 대한 다양한 지표
제공



■ 시사점 2) 인스타그램에 맞춤화된 지표 제공

- ✓ 기존 팔로워 수, 좋아요 수, 전체 팔로워 수에서만 판단하
는 지표에서 최근 게시물의 업로드 정도의 횟수, 광고 게
시물에 대한 노출 정도 등에 대한 지표를 제공

Expertise(10 Feature)

Trustworthiness(1 Feature)

Attractiveness(3 Feature)

실정에 맞는
지표 참고

■ 시사점 3) 나노인플루언서 새로운 판단 기준 제공 가능

- ✓ 팔로워 수가 적더라도 효과적인 인플루언서로 자리잡을
수 있는 나노 인플루언서에 대한 새로운 판단 기준 제공



데이터마이닝 기법을 통해 본 과제를 진행할 때에 크게 데이터 수집이 어렵고 선행 연구가 제한적이며, 연구 대상의 분야가 한정 되어있다는 한계점이 존재했습니다

데이터 수집	특정 분야 한정	참고 자료 부족
<p>현실적 한계</p> <ul style="list-style-type: none"> ✓ 노트북을 통한 데이터 수집에는 컴퓨터의 속도에 대한 문제가 있어 팔로워 목록을 불러오는 데 문제가 생깁니다. ✓ 게시물 하나하나 확인하고 들어가며 데이터를 수집하는 인스타그램 특성상 데이터를 모으는데 많은 시간이 걸립니다. 	<p>패션/뷰티 분야 한정</p> <ul style="list-style-type: none"> ✓ 위 연구는 패션/뷰티 인플루언서에 대해 한정을 지어 진행을 했기 때문에, 다른 분야(ex. 운동, 음식)에 대한 분석 결과는 다를 수 있습니다. ✓ 특정 분야에 대해 진행하다 보니, 수집 할 수 있는 대상이 한정이 되어있었습니다. 	<p>인스타그램 분석 논문 수 부족</p> <ul style="list-style-type: none"> ✓ 인스타그램에 존재하는 인플루언서에 대한 연구한 자료가 거의 없었기 때문에 새롭게 측정할 지표에 대해 조사하고만 들어 나갔습니다. ✓ 참고한 논문은 1990년 논문이기 때문에 최근 온라인과 모바일을 통해 뜨고 있는 Celebrity에 대해 새롭게 정의하고 있는 논문이 부족했습니다.
<p>인스타그램 정책</p> <ul style="list-style-type: none"> ✓ 인스타그램 정책 상 특정 게시물 수 이상을 불러오지 못하게 되어 있습니다. ✓ 인스타그램 특성상 로그인을 한 후 접근을 해야 하는데, 여러 번 사용 후엔 가계정의 비밀번호를 바꾸거나, 다시 가계정에 대한 인증을 해야하는 등의 상황이 생깁니다. 		

References

- 1) Cassandra Schwartz. (2017). *2017 Social Media Industry Benchmark Report*. Retrieved from <https://www.rivaliq.com/blog/2017-social-media-industry-benchmark-report/>
- 2) John Analyst. (2020). Retrieved from <https://john-analyst.medium.com/smote%EB%A1%9C-%EB%8D%B0%EC%9D%B4%ED%84%B0-%EB%B6%88%EA%B7%A0%ED%98%95-%ED%95%B4%EA%B2%B0%ED%95%98%EA%B8%B0-5ab674ef0b32>
- 3) Roobina Ohanian(1990), Construction and Validation of a scale to Measure Celebrity Endorsers' Perceived Expertise, Trustworthiness, and Attractiveness , Journal of Advertising Vol.19, p39–52.
- 4) 김희숙. (2020). 데이터 마이닝을 이용한 유튜브 인기 동영상 콘텐츠 분석. 한국디지털콘텐츠학회 논문지, 21(4), 673-681.
- 5) 남연주, 김용호(2021). 인스타그램 인플루언서 마케팅의 팔로워 지각 효과에 관한 연구, 미디어, 젠더&학회 Vol.36, no1, p279-310.
- 6) 문지원, 김원경. (2020). 브랜드 커뮤니케이션 활성화를 위한 효과적인 인플루언서(Influrencer) 마케팅 전략 개발 제안, 21(1), 197-210.
- 7) 한국무역신문. (2021). “팔로워 충성도 높은 ‘나노 인플루언서’를 주목하라”. Retrieved from <https://www.kita.net/cmmrcInfo/cmmrcNews/cmmrcNewsDetail.do?pageIndex=1&nIndex=61977&sSiteid=1>
- 8) 국내 X 쇼핑 어플리케이션 브랜드 영업관리/마케팅 담당자. (2021년 05월 15일). 화상 인터뷰.
- 9) 우현우. (2021) 2021-1 데이터마이닝 이론 및 응용 실습자료.

감사합니다