AI 기반의 온라인 쇼핑몰 고객 리뷰 데이터를 통한 상품 인사이트 도출 및 시각화

Deriving and Visualizing Product Insights from Shopping Mall Customer Review Data Using AI

김 원 명*, 송 한 춘**

Kim Won Myeong*, Song Han Chun****

sanho@seoil.ac.kr

Abstract

This paper presents the development of a program that analyzes customer review data from online shopping malls to derive product insights. The process of deriving product insights from customer review data for specific products in an online shopping mall involves several steps: web crawling, data preprocessing, sentiment analysis, topic modeling. we analyze customer reviews of specific products in shopping malls to derive insights and implement visualizations to make this information easily understandable for customers. The extraction and visualization of insights for specific products in online shopping malls can aid consumers in their purchasing decisions.

keyword : product insights, sentiment analysis, topic modeling, data visualization

I. 서 론

인터넷 쇼핑몰에서 소비자 리뷰는 소비자의 구매 결정에 큰 영향력을 행사하는 지표이다. 온라인 쇼핑 기업들은 소비자 리뷰 데이터를 수집하기 위하여 다양한 프로모션 행사를 진행하고 있다.[1] 온라인 시장조사 전문기업인 엠브레인 트렌드 모니터가 전국 만 19~49세 성인 남녀 1,200명을 대상으로 한 소비자 리뷰와 관련한 설문 조사에서 소비자 리뷰 영향력이 매우 큰 것으로 조사되었다.[2] 소비자 리뷰를 기반으로 객관적인 입장에서 제품에 대한 의견을 파악할 수 있는 프로그램의 개발이 필요하다.

본 논문에서는 'AI 허브 속성 기반 감정 분석 데이터'와 온라인 쇼핑몰 특정 상품 소비자 리뷰 데이터를 크롤링하여 수집한 후에, 이들 데이터를 NLP(Natural Language Process) 감정 분석과 LDA(Latent Dirichlet Allocation) 토픽 모델링을 사용하여 제품에 인사이트를 도출하는 프로그램과 이를 쉽게 확인할 수 있도록 시각화 하는 프로그램을 개발하였다.

*서일대학교 정보통신공학과 학생

**서일대학교 정보통신공학과 교수 <교신저자>

II. 관 련 연 구

2.1 오피니언 마이닝

오피니언 마이닝은 텍스트 형태의 데이터를 분석하여 특정 현상에 대한 의견을 분석하여 유용한 정보로 재가공하는 기술이다.[3] 오피니언 마이닝은 텍스트 문서 수집, 감성과 관리 없는 부분을 제거하는 주관성 탐지, 단어의 극성을 분석한다. 최근 AI 기술 발전과 함께 오피니언 마이닝을 활용하여 제품이나 서비스에 대한 소비자 평가를 자연어 처리를 통해 상품, 서비스, 기업에 대한 인사이트를 도출한다[3].

2.2 토픽 모델링

구조화되지 않은 대량의 텍스트를 분석하여 숨겨져 있는 주제 구조를 발견하고 범주화하는 통계적 추론 알고리즘이다. 토픽 모델링의 대표적인 알고리즘은 LSA(Latent Semantic Analysis)와 LDA(Latent Dirichlet Allocation) 방식이다. LSA는 SVD 특이값 분해를 활용해서 중요한 토픽만 남기고 불필요한 토픽은 제거하는 기법이다. LDA 기법은 단어가 특정 토픽에 존재할 확률과 문서에 특정 토픽이 존재할 확률을 결합 확률로 추정하여 토픽을 추출하는 기법이다.[4]

2.3 지도학습

지도학습은 훈련 데이터(Training Data)로부터 하나의 함수를 유추해내기 위한 기계 학습(Machine Learning)의 방법이다. 입력 데이터를 통해 머신러닝 모델이 예측한 데이터와 정답 데이터를 비교하면서 오차를 줄인다.[5]

본 논문에서는 4가지 지도학습 분류 알고리즘 로지스틱 회귀 분류, 베르누이 나이브 베이즈, 커널 서포트 벡터 머신, 랜덤 포레스트를 사용한다. 로지스틱 회귀 방식은 데이터가 어떤 범주에 속할 확률을 0에서 1 사이의 값으로 예측하고 확률이 더 높은 범주로 분류해주는 알고리즘이다.[6] 베르누이 나이브 베이즈는 새로운 정보를 토대로 어떤 사건이 발생했다는 주장에 대한 신뢰도를 갱신해 나가는 나이브 베이즈 분류기의 한 종류로 고차원 데이터에서 잘 작동하여 텍스트 데이터에 효과적이다.[7] 커널 서포트 벡터 머신은 데이터를 가장 잘 분리할 수 있는 최적의 초평면을 찾아내는 알고리즘이다. 입력 데이터에서 단순 초평면으로 정의되지 않는 더 복잡한 모델을 만들 수 있도록 확장한다[8]. 랜덤 포레스트는 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성한 다수의 결정 트리로부터 분류한다[9].

2.3 데이터 시각화

데이터 시각화(Data Visualization)는 수집된 데이터를 그래픽 형식으로 표현하여 새로운 통찰력을 제공하거나 정보를 보다 명확하고 효과적으로 전달한다. 이는 데이터의 의미를 직관적으로 파악하고 의사소통을 하는 데 도움을 준다. 데이터의 시각화는 마인드 맵, 뉴스 표현, 관계 표현, 데이터 표현, 웹사이트 표현 등 다양한 분야에 활용 수 있다.[10]

Ⅲ. 시스템 설계 및 구현

3.1 개발 시스템 설계

본 논문에서는 Colab(2024-05-22) 분석 환경에서 Python (3.11.8) 언어를 사용하여 프로그램을 개발하고 데이터 저장은 Microsoft Excel 2016을 오피스 소프트웨어를 사용한다.

프로그램 개발의 순서는 “AI 허브”와 “O”사 온라인 상품 리뷰 데이터를 수집 후 긍정과 부정으로 수집 데이터를 분류한다. 데이터 품질을 높이고 분석 결과의 정확성을 높이는 데이터 전처리를 실행한다. 전처리한 데이터를 여러 지도학습 알고리즘 모델에 학습시켜 일반화 성능이 가장 좋은 알고리즘 모델을 사용하여 목표 데이터를 긍정과 부정으로 분류한다. 분류한 데이터는 토픽 모델링을 활용하여 시각화한다. 아래 <그림 1>은 전체 시스템의 흐름도이다.

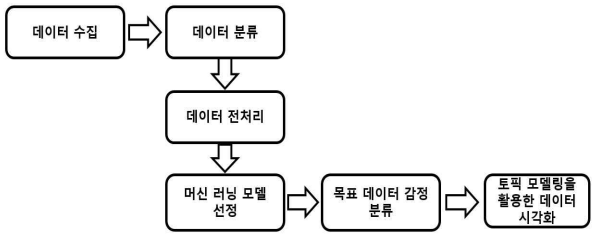

그림 1. 시스템 흐름도

3.2 데이터 수집 및 분류

모델 학습을 위한 훈련 데이터는 “AI 허브 속성 기반 감정 분석 데이터”의 쇼핑몰 화장품 리뷰 데이터를 활용한다. 지도학습을 위한 데이터 감정 분류는 “AI 허브 속성 기반 감정 분석 데이터”에서 제공한 긍정과 부정 라벨링을 기준으로 분류한다.

모델 검증을 위한 테스트 데이터는 온라인 최대 코스메틱 쇼핑몰 “올리브영” 인기 상품 A사의 “마스크 팩” 제품의 소비자 리뷰 데이터를 크롤링하여 csv 파일 형식으로 저장한다. 크롤링한 테스트 데이터는 소비자 리뷰 별점 5점과 4점을 긍정 데이터, 별점 3점을 중간 데이터, 별점 2점과 1점을 부정적 데이터로 분류한다. 본 논문에서는 소비자 평점 3점인 중간 데이터를 감정 분류 모델을 활용해서 긍정/부정으로 분류할 것이기 때문에 “목표 데이터”라 지칭할 것이다.

	review	label
0	유통 기한 넉넉하고	1
1	구성 많아서 선물 하기 좋네요	1
2	구성 알차고	1
3	촉촉하고 너무 좋아용	1
4	대 용량 넉넉하게 사용 할 수 있고	1

그림 2. 훈련 데이터

userId	rating	reviews
와바밤	1	사실 마스크 팩 한 달동 안 사용 것 아니라 한 달 리뷰 하기 얘 매하 긴한데 어쨌...
와바밤	1	배송비 가격 맞출 때 유용하게 매번 구매 있습니다 자극 없이 순해여
보부상출신	1	나이 먹으니까 피부 고민 점차 주름 미백 쪽 빠지는데데 일리 피부 결 톤 케어 하기...
파이지니	1	1일 1 팩 하기 좋은 가성 비 마스크 팩 메디 힐 마스크 팩 너무 유명하죠...
보부상출신	1	요즘 제 피부 상태 탄력 지금 저 세상 가있는데 복구 위해 콜라겐 마스크 팩 이용 ...

그림 3. 테스트 데이터

3.3 데이터 전처리

수집한 데이터는 한국어 기반 데이터이므로 Okt 형태소 분석기를 사용하여 조사, 구두점, 알파벳 형태소를 제거한다. Okt 클래스를 상속받아 __getstate__와 __setstate__ 함수를 오버라이딩하여 Okt 객체를 직렬화와 역 직렬화하면 학습된 모델과 파이프라인을 저장하고 재사용할 수 있다.

```
class PicklableOkt(Okt):
    def __init__(self, *args):
        self.args = args
        Okt.__init__(self, *args)

    def __getstate__(self):
        return {'args': self.args}

    def __setstate__(self, state):
        self.__init__(*state['args'])
```

그림 4. 데이터 전처리 코드

3.4 머신러닝 알고리즘별 최적화 모델 실험 및 결과 데이터를 활용한 분석 모델 선정

3.4.1 머신러닝 알고리즘별 최적화 모델 실험 과정

지도학습 데이터 분류에 사용되는 로지스틱 회귀 분류, 베르누이 나이브 베이즈, 랜덤 포레스트, 서포트 벡터 머신을 사용하여 알고리즘별 분류 성능을 실험한다.

본 논문에서는 파이프라인과 그리드 서치를 사용하여 TF-IDF 통계적 기법 매개변수 gram_range, min_df와 지도학습 알고리즘별 매개변수 alpha, C, gamma, max_depth, max_features, n_estimators의 값의 변화에 따른 모델 성능을 비교하여 알고리즘별 최적화 파라미터 값을 찾는다. <표 1>은 지도학습 알고리즘별 사용한 매개변수를 정리한 표이다. <그림 5>는 4개의 머신러닝 알고리즘 중 SVM 알고리즘의 파라미터 “C”와 “gamma”의 값과 TF-IDF의 “min_df”와 “ngram_range”를 변경하여 훈련 데이터에 대한 최적 파라미터를 탐색하는 코드이다. 나머지 세 개의 알고리즘도 <그림 5> 코드와 유사하게 파이프라인과 그리드 서치를 사용하여 모델별 최적 파라미터 값을 찾는다.

표 1. 지도학습 알고리즘 매개변수

알고리즘	매개변수
선형 회귀 분류	C
BernoulliNB	alpha
SVM	C, gamma
RandomForest	max_depth, max_features, n_estimators

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.pipeline import make_pipeline
from sklearn.model_selection import GridSearchCV

# 직렬화 가능한 PickableOkt 사용하여 파이프라인 생성
param_grid = {
    'tfidfvectorizer__min_df': [1,3,5],
    'tfidfvectorizer__ngram_range': [(1, 1), (1, 2), (1, 3)],
    'svc__C': [0.1, 1, 10],
    'svc__gamma': [0.1, 0.01, 0.001]
}

pipe = make_pipeline(TfidfVectorizer(tokenizer=PicklableOkt().morphs), SVC(kernel='rbf'))
grid = GridSearchCV(pipe, param_grid, n_jobs=-1)
grid.fit(text_test, y_test)
```

그림 5. 최적 파라미터 탐색 코드

3.4.2 실험 결과 데이터를 활용한 분석 모델 선정

선형 분류 알고리즘은 “tfidfvectorizer” 벡터화 매개변수인 “min_df”, “ngram_range”를 3과 (1, 3)으로 설정하고 선형 회귀 분류 매개변수 “C”값을 10인 선형 분류 알고리즘 모델을 사용하면 훈련 세트 정확도 89%와 테스트 세트 정확도 77%를 기록하며 최적의 성능 결과가 나타난다. BernoulliNB 알고리즘은 “tfidfvectorizer” 벡터화 매개변수인 “min_df”, “ngram_range”를 3과 (1, 2)으로 설정하고 BernoulliNB 매개변수 “alpha”값을 0.1로 설정하면 훈련 세트 정확도 88%와 테스트 세트 정확도 77%를 기록하며 최적의 성능 결과가 나타난다. SVM 알고리즘은 “tfidfvectorizer” 벡터화 매개변수인 “min_df”, “ngram_range”를 1과 (1, 2)로 설정하고 SVM 알고리즘 매개변수 “C”를 0.1로 설정하면 훈련 세트 정확도 91%와 테스트 세트 정확도 77%를 기록하며 최적의 성능 결과가 나타난다. 마지막으로 랜덤 포레스트 알고리즘은 “tfidfvectorizer” 벡터화 매개변수인 “min_df”, “ngram_range”를 1과 (1, 3)으로 설정하고 랜덤 포레스트 알고리즘 매개변수 “max_depth”“max_features”,“n_estimator”를 20, “sqrt”, 400으로 설정하면 훈련 세트 정확도 81%와 테스트 세트 정확도 73%를 기록하며 최적의 성능 결과가 나타난다.

<그림 6>에서의 모델 과대 적합 현상은 <그림 7> 테스트 데이터 별점과 리뷰 내용이 상반된 의미적 오류로 인한 것으로 예상된다. 테스트 데이터의 오류를 수정하면 과대 적합 현상이 해소되어 훈련 세트 성능이 가장 높은 SVM 알고리즘 모델의 테스트 세트 정확도가 실험 모델들 중 가장 높을 것으로 예상된다. 따라서 본 논문에서는 SVM 모델을 사용하여 목표 데이터를 긍정과 부정으로 감정 분류한다.

알고리즘 \ 성능	훈련 세트 정확도	테스트 세트 정확도	최적 파라미터
LogisticRegression	0.89	0.77	'logisticregression__C': 10, 'tfidfvectorizer__min_df': 3, 'tfidfvectorizer__ngram_range': (1, 3)
BernoulliNB	0.88	0.77	'bernoullinb__alpha': 0.1, 'tfidfvectorizer__min_df': 3, 'tfidfvectorizer__ngram_range': (1, 2)
SVM	0.91	0.77	'svc__C': 10, 'svc__gamma': 0.1, 'tfidfvectorizer__min_df': 1, 'tfidfvectorizer__ngram_range': (1, 2)
RandomForest	0.81	0.73	'max_depth': 20, 'randomforestclassifier__max_features': 'sqrt', 'randomforestclassifier__n_estimators': 400, 'tfidfvectorizer__min_df': 1, 'tfidfvectorizer__ngram_range': (1, 3)

그림 6. 지도학습 알고리즘 모델 성능 결과

chris1****	0	너무 너무 좋아요 강 추합니다 최고 재구 매각 입니다
라비링스	0	한번 사용 해서 다음 날 확실히 피부 달라 진정 너무 좋아요 ㅎㅎ
pag****	0	베리 굿굿 양도 많고 진정 용 딱 좋습니다 앞 평생 장 착 할 듯

(참조. 0 => 부정, 1=> 긍정)

그림 7. 테스트 세트 데이터 오류

3.5 목표 데이터 감정 분석

<그림 8> 그래프는 300개의 목표 데이터를 감정 분류한 분류 모델의 예측값에 대한 비율이다. SVM 모델은 300개의 목표 데이터 중 158개의 데이터를 긍정적 성격 리뷰로 분류하고 나머지 142개의 데이터를 부정적 성격 리뷰로 분류한다.

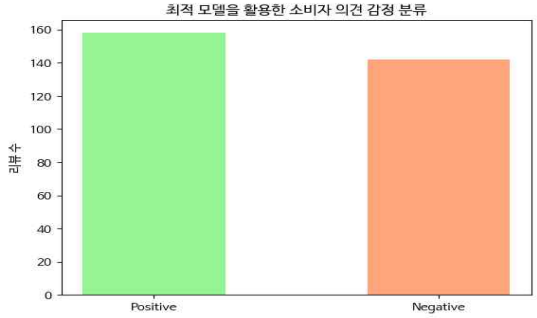

그림 8. 소비자 리뷰 감정 분석

4. 도출한 인사이트 결과의 시각화

4.1 시각화를 위한 데이터 프레임 결합

<그림 8> 순서도와 같이 목표 데이터를 SVM 알고리즘을 활용해 감정 분류한 긍정적 소비자 의견 데이터인 “olive_med_labeled_pos_df”를 소비자 평점 4점, 5점 데이터 프레임“olive_pos_df”와 병합시켜 “final_pos_df” 긍정 데이터 프레임을 생성하고 감정 분류한 데이터 중 부정적 소비자 의견인 “olive_med_labeled_neg_df”를 소비자 평점 1점, 2점 데이터 프레임이 “olive_neg_df”에 병합하여 부정 데이터 프레임 “final_neg_df”를 생성한다.

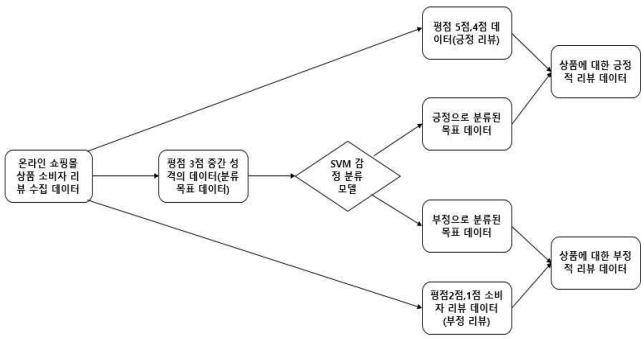

그림 8. 감정별 소비자 리뷰 데이터 결합 순서도

4.2 토픽 모델링 시각화

4.2.1 토픽 모델링 시각화 대상과 형식

본 논문에서는 LDA 토픽 모델링 방식으로 상품에 대한 감정별 소비자 리뷰 데이터를 분석하여 5개의 주제를 선정하고 주제별 연관성과 핵심어를 Intertopic Distance Map과 막대그래프 형식으로 시각화하였다.

.

4.2.2 토픽 모델링 시각화 과정

<그림 9>와 같이 “final_pos_df”와 “final_neg_df” 데이터 프레임 “review” 열의 소비자 리뷰 텍스트 데이터를 corpora.Dictionary를 사용하여 단어 사전 생성한 후 doc2bow 라이브러리를 활용해 말뭉치를 생성한다. Gensim의 LDA 모델의 파라미터 값인 num_topic (생성 토픽의 개수)과 passes(전체 데이터 세트 반복 횟수)를 각각 5와 10으로 설정한다. lda_visualize 함수에 model, corpus, dictionary 변수를 대입해서 pyLDAvis 라이브러리를 사용해서 시각화한다.

```python
def lda_modeling(review_prep):
    # 사전 구축
    dictionary = corpora.Dictionary(review_prep)

    # 말뭉치 생성
    corpus = [dictionary.doc2bow(review) for review in review_prep]

    # LDA 모델 학습
    NUM_TOPICS = 5 # 토픽의 수
    PASSES = 10 # 알고리즘이 전체 데이터 세트를 반복하는 횟수
    model = gensim.models.ldamodel.LdaModel(corpus,
                                            num_topics=NUM_TOPICS,
                                            id2word=dictionary,
                                            passes=PASSES)
    return model, corpus, dictionary
```

```python
def lda_visualize(model, corpus, dictionary, RATING):
    pyLDAvis.enable_notebook()
    result_visualized = pyLDAvis.gensim_models.prepare(model, corpus, dictionary)
    pyLDAvis.display(result_visualized)
    # 시각화 결과 저장
    RESULT_FILE = 'lda_result_' + RATING  +'.html'
    pyLDAvis.save_html(result_visualized, RESULT_FILE)
```

그림 9. LDA 토픽 모델링과 시각화 코드

4.2.3 토픽 모델링 시각화 그래프 분석과 인사이트 도출

1) 시각화 그래프 특징을 활용한 그래프 분석

Intertopic Distance Map에서의 원들은 서로 다른 주제를 의미하고 원의 넓이는 분석 데이터 안에서의 해당 주제를 구성하는 키워드들의 빈도수를 의미한다.

그림 9. 긍정적 소비자 리뷰 시각화 그래프

Intertopic Distance Map에서의 원들 간의 거리는 주제의 유사도를 의미한다. 막대그래프는 분석 데이터 안에서 해당 주제별 키워드 빈도수를 나타낸다. <그림 9> 막대그래프에서 해당 주제별 키워드를 확인하면 1번, 2번 주제 모두 1순위로 '진정' 키워드의 빈도수 수치가 가장 높다. 하지만 1번 주제에서 2순위와 3순위 키워드가 '시트'와 '효과'인 반면 2번 주제에서 2순위와 3순위의 키워드는 '에센스'와 '티트리'로 다소 의미적 차이가 있다. 따라서 1번 원과 2번 원이 모두 Intertopic Distance Map 그래프의 2사분면에 위치하지만 겹치지는 않는다.

<그림 10> 부정적 소비자 리뷰 시각화 막대그래프에서 1번, 2번, 3번 주제를 구성하는 키워드를 보면 1번, 2번, 3번 주제 모두 '진정'과 '티트리'를 키워드로 포함하고 있어 Intertopic Distance Map 그래프 1사분면에 3개의 원이 모여있다.

그림 10. 부정적 소비자 리뷰 시각화 그래프

2) 시각화 그래프 결과 인사이트 도출

시각화 그래프를 분석하면 소비자들은 A사 제품에 긍정적인 평가를 한 소비자들은 '진정', '시트', '보습', '가격' 부분에 긍정적이나 부정적인 평가를 한 소비자들은 '여드름', '티트리', '리뉴얼', '자극', '진정'에 대해 부정적 인식을 가지고 있다는 감정별 상품 인사이트를 도출할 수 있다.

Ⅳ. 결 론

최근 이커머스 플랫폼과 SNS 마케팅으로 온라인 소비 시장에서 거래량이 증가하면서 소비자들은 상품에 대한 다양한 의견을 자유롭게 공유한다. 본 논문에서 개발한 프로그램은 특정 상품에 대한 소비자들의 다양한 의견을 수집하고 분석하여 상품에 대한 인사이트를 제공한다. 소비자는 인사이트를 활용하여 상품에 대한 많은 양의 소비자 의견을 간결하게 확인할 수 있어 상품 구매 의사결정 시간을 단축한다.

본 논문에서는 'AI 허브 속성 기반 감정 분석 데이터'와 온라인 쇼핑몰 특정 상품 소비자 리뷰 데이터를 크롤링하여 수집한 후에, 이들 데이터를 NLP(Natural Language Process) 감정 분석과 LDA(Latent Dirichlet Allocation) 토픽 모델링을 통하여 특정 상품의 인사이트를 도출하는 프로그램을 개발하였고, 도출된 결과를 쉽게 확인할 수 있도록 시각화 하는 프로그램을 개발하였다.

본 논문에서 개발한 온라인 쇼핑몰에서 특정 제품에 대한 통찰을 추출하고 시각화하는 프로그램은 온라인 쇼핑몰에서 소비자의 구매 결정에 도움을 줄 수 있을 것이다.

참 고 문 헌

[1] 이상미 기자 '호갱'이 되고 싶지 않은 소비자, '소비자 리뷰' 확인, 산업일보(https://kidd.co.kr/news/196005)

[2] 신경식, '구매 후기 한 줄에 고객의 이런 속마음이 마케팅 난제, 속 시원히 풀어주는 분석(DBR 261호, 2018.11)

[3] 윤병운, ;인공지능을 활용한 오피니언 마이닝 - 소셜 오피니언 마이닝은 무엇인가? (Samsung SDS, 2017.05)

[4] 유원준 외1, '딥 러닝을 이용한 자연어 처리 입문' (e-book, 2024.04)

[5] Bebsae, '선형 분류와 선형 회귀', 2021.11 (https://dev-ryuon.tistory.com/65)

[6] 오윤태, '로지스틱 회귀와 서포트 벡터 머신러닝 성능 비교' (국민대학교 일반대학원 학위논문(석사), 2019)

[7] Wikipedia, '나이브 베이즈 분류' (https://ko.wikipedia.org/wiki/skdlqm 베이즈 분류)

[8] 데이터 시각화란 무엇인가요? (https://aws.amazon.com/ko/what-is/data-visualization)