구글 앱스토어 리뷰 감성 분석 및 토픽 모델

- 구글 앱스토어 사용자 리뷰 데이터를 활용하여 텍스트 분석-

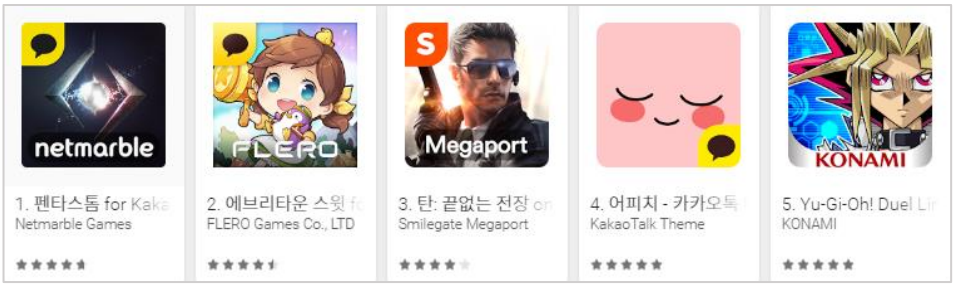
학번	학과	이름
182STG12	통계학과	오혜윤

목차

1. 배경

1.1 앱 스토어 산업 성장 및 현황

현재 구글 플레이 스토어에는 끊임없이 어플이 출시되고 있다. 사용자들은 다양한 어플을 활용한 후 리뷰를 남기고 해당 어플 개발자와 소통한다. 또한, 어플은 끊임없이 사용자의 리뷰를 통해 개선되며, 그에 따라 평점과 순위가 정해진다. 빠르게 변화하는 앱 스토어 시장에서 인기있는 어플로 자리잡기 위해서는 사용자의 리뷰가 절대적으로 중요한 것이다.



2008 년 500 개의 어플리케이션(앱)으로 시작한 애플 앱스토어는 2017 년 1 월에는 220 만 개가 넘는 앱을 유통시키고 있다. 한편 애플이 앱스토어를 출시한 후 석달 뒤 안드로이드 마켓이 세상에 모습을 드러냈다. 이후 2012 년 3 월 구글플레이(Google Play)로 이름을 바꾸며 애플 앱스토어와 모바일 앱 생태계의 양대 산맥을 이루고 있다. 앱 숫자는 2017 년 초 애플 앱스토어를 능가하는 270 여만 개에 이른다.

누적 다운로드 수는 애플 앱스토어의 경우 2017 년 6 월까지 1,800 억회, 구글 플레이의 경우 2016 년까지 약 650 억회이다. 2016 년까지 누적 다운로드 수는 애플 앱스토어가 구글 플레이를 앞서고 있지만, 최근 수년간 구글 플레이에서의 다운로드 수가 애플 앱스토어를 누르고 있다.

최근 3년 구글 플레이, 애플 앱스토어 앱 다운로드 수 (출처: App Annie)

[1]

그림 1: 앱스토어 다운로드수

[1] 모바일 앱 생태계 현황과 인스턴트 앱의 시사점 (http://slownews.kr/67025)

스마트폰 사용자들은 월 평균 30 개의 앱을 이용하고 있다고 한다. 이는 스마트폰에 설치되어 있는 전체 앱의 약 삼분의 일에서 반 정도 수준이다. 물론 국가별로 편차가 있는데, 사용량이 많은 주요 국가들의 월 사용 앱 숫자와 설치된 앱 숫자는 아래 도표와 같다.

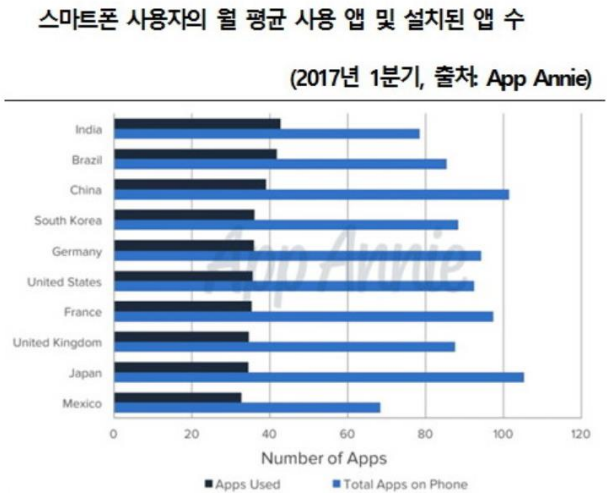


그림 2: 스마트폰 사용자의 설치 앱 수

따라서, 많은 스마트폰 사용자들은 점점 매우 많은 앱들을 사용하고 있다. 특히, 앱스토어 시장에서도 구글 앱스토어의 인기가 커진 만큼 구글 앱스토어 시장 분석이 중요할 것이라 예상한다.

1.2 주제 선정 이유

865 개의 어플에 대한 리뷰 감성 분석을 통해 우리는 해당 어플 사용자의 태도, 의견, 성향을 알아볼 수 있다. 리뷰 데이터의 긍정, 부정, 중립을 판단하여 어떠한 어플이 인기 있는지 구분할 수 있으며, 해당 어플이 흥행할 것인가에 대해서도 알아볼 수 있다.

또한, LDA 토픽 모델을 활용하여 각각의 리뷰에 대한 토픽을 추출할 수 있다. 리뷰 토픽 분석을 통하여 우리는 해당 어플의 가장 큰 특징을 살펴볼 수 있다. 즉, 어플을 사용하고 있는 사용자의 불편한 점이나 좋은 점의 추세와 중요도를 파악할 수 있는 것이다.

감성분석과 LDA 토픽 모델을 통해, 급변하는 어플 시장의 추세를 살펴볼 수 있으며, 이러한 결과를 통해 어플 개발자는 사용자의 입장을 고려한 편리한 서비스를 제공할 수 있을 것이다.

[2] 모바일 앱 생태계 현황과 인스턴트 앱의 시사점 (http://slownews.kr/67025)

1.3 분석 방법론

*감성 분석

각각의 리뷰 데이터의 속성을 추출하여, 해당 속성에 대한 감성의 극성을 분류한다. 모델링 방법으로 Naïve Bayes, SVM, Logistic, RandomForest 를 사용하고, 각각의 모델의 정확도를 비교한다. 이를 통해, 우리는 구글 앱스토어 리뷰 데이터의 감성을 가장 잘 분류하는 모델을 찾아낼 수 있다. 이렇게 도출된 결과는 향후 앱 스토어 시장의 방향성을 제시해줄 것이다.

*LDA 토픽 모델

리뷰 데이터에 내재된 토픽들을 추출하여 감추어진 화제 혹은 정리된 개념을 산출한다. LDA 토픽 모델을 통해 특정 토픽에 특정 단어가 나타날 확률을 제시하거나 각 리뷰 데이터의 토픽 분포를 제시한다. 이러한 텍스트 데이터의 토픽을 추출함을 통해 해당 어플 사용자의 관심사나 여론을 분석할 수 있을 것이다.

2. 데이터 전처리

2.1 데이터 설명

Kaggle에서 제공하는 google app store 데이터는 각각의 어플에 대한 정보(store) 와 사용자 리뷰(user) 데이터로 이루어져 있으며, 텍스트 분석을 위해 **사용자 리뷰(user) 데이터**만을 사용하였다.

1) 사용자 리뷰 전체 데이터는 총 64295개로 많은 결측치를 포함한다.
2) 1074개의 unique한 어플로 이루어져 있으며, 총 5개의 column을 가진다.
3) 각 어플에 대한 리뷰와 리뷰 감성 column들로 이루어져 있다.

전체 데이터 예시

App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
10 Best Foods for You	NaN	NaN	NaN	NaN
10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000
10 Best Foods for You	Best idea us	Positive	1.00	0.300000

2.2 데이터 전처리

2.2.1 결측치 처리 및 컬럼 조정

1) 전체 사용자 리뷰 데이터셋은 26971개의 결측치를 포함한다. 따라서, 전체 데이터셋에서 결측치를 제거하여 남은 총 37324개의 데이터를 사용하였다.
2) 전체 데이터셋에서 필요 없는 컬럼들은 제거하고 어플 이름 "APP"과 감성 라벨 "Sentiment" 컬럼 두 가지만을 취하였다.
3) "Sentiment" 컬럼에는 각각의 리뷰에 대한 Positive, Negative, Neutral 총 3가지 감성으로 분류된 라벨이 존재한다. 따라서, 이를 이용하여 감성 분류 분석 모델링을 행하였다.
4) "Sentiment" 컬럼의 Positive, Negative, Neutral은 각각 0,1,2로 encoding하여 진행하였다.

최종 데이터 예시

	App	Translated_Review	Sentiment
64222	Housing-Real Estate & Property	Most ads older many agents ..not much owner po...	Positive
64223	Housing-Real Estate & Property	If photos posted portal load, fit purpose. I'm...	Positive
64226	Housing-Real Estate & Property	Dumb app, I wanted post property rent give opt...	Negative
64227	Housing-Real Estate & Property	I property business got link SMS happy perform...	Positive
64230	Housing-Real Estate & Property	Useless app, I searched flats kondapur, Hydera...	Negative

2.2.2 데이터 정제

감성 분석을 행하기에 앞서 텍스트 데이터를 정제해주었다.
먼저, ":) # $ @ ()!-/*" 등과 같은 기호를 제거하고, 모두 소문자로 바꾸어 주었다.

텍스트 변환 예시

Original text -> Manipulated text	
I like eat delicious food. That's I'm cooking food myself, case "10 Best Foods" helps lot, also "Best Before (Shelf Life)"	i like eat delicious food that s i m cooking food myself case best foods helps lot also best before shelf life

이후, 토큰화 작업을 실시하였다. 그리고 lemmatization을 통해 어간만을 추출하여 복수형 등을

수정하였다.

Lemmatization 예시

Manipulated text
i like eat delicious food that s i m cooking food myself case best food help lot also best before shelf life

마지막으로, i, am, me 등과 같은 모든 불용어를 제거해주었다.

불용어 제거 예시

Manipulated text
like eat delicious food cooking food case best food help lot also best shelf life

3. 모델링

3.1 EDA

3.1.1 Plot

3.1.1.1 Sentiment Count Plot

Sentiment 분포를 알아보기 위해 그래프를 그려보았다.
Positive : 0, Negative : 1 , Neutral : 2라고 할 때, Positive의 비율이 압도적으로 높았다.

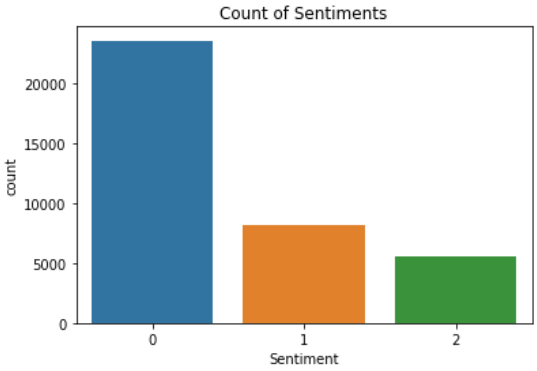


그림 3 : Sentiment Count 분포

3.1.1.2 Most Common Words Plot

가장 많이 사용된 단어를 확인하기 위하여 count 를 구한 후, plot 을 그려보았다.

game 이 10374 번으로 가장 많이 사용된 단어로 나타났으며, 압도적으로 빈도수가 높음을 확인할 수 있었다. 그 다음은 time, like 순으로 나타났다.

Most Common Words Count

Number of words: 17580

Most common words: [('game', 10374), ('time', 5658), ('like', 5595), ('good', 5294), ('get', 5153), ('app', 5025), ('great', 4819), ('love', 4783), ('ad', 3467), ('work', 3452), ('would', 3417), ('really', 3096), ('update', 3047), ('even', 2889), ('make', 2861)]

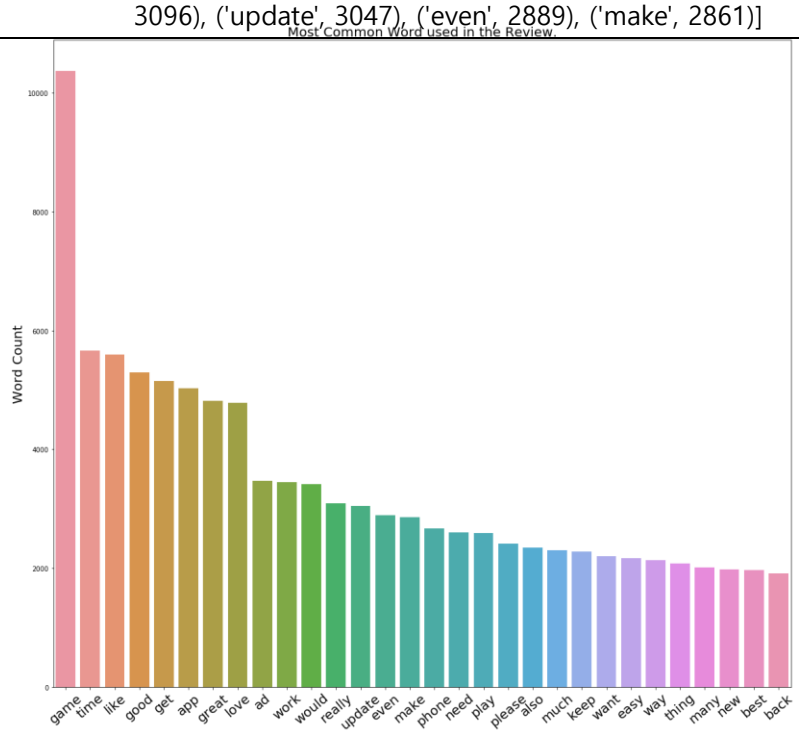


그림 4: Most Common Words Plot

가장 많이 사용된 단어의 긍정, 부정, 중립 감성 각각의 Count 를 보기 위하여 plot 을 그려보았다.

그림 5: 긍정, 부정, 중립 별 Most Common Words Plot

가장 많이 사용되었던 단어 game 은 긍정, 부정에서 모두 많이 나타났지만, 중립에서는 별로 나타나지 않았다. 반면, time 은 모든 감성에서 높은 빈도수를 나타냈다. 그리고 보통 긍정 단어로 분류되는 good, great 은 예상했던 것과 같이 긍정에서 가장 높은 빈도수를 나타냈다. 하지만, 중립 보다는 부정 리뷰에서 더 많은 빈도수를 보였다.

3.1.2 워드 클라우드

워드 클라우드를 그려본 결과, 긍정 리뷰에 대해서는 good, love, great 이 가장 많이 나타났다. 반면, 부정 리뷰에 대해서는 game, time, like, get 이 많이 나타났다. 또한, 중립 리뷰에 대해서는 work, like, time 이 가장 많이 나타났다.

그림 6: 긍정, 부정, 중립 별 WordCloud

3.2 감성분석

3.2.1 모델링 과정

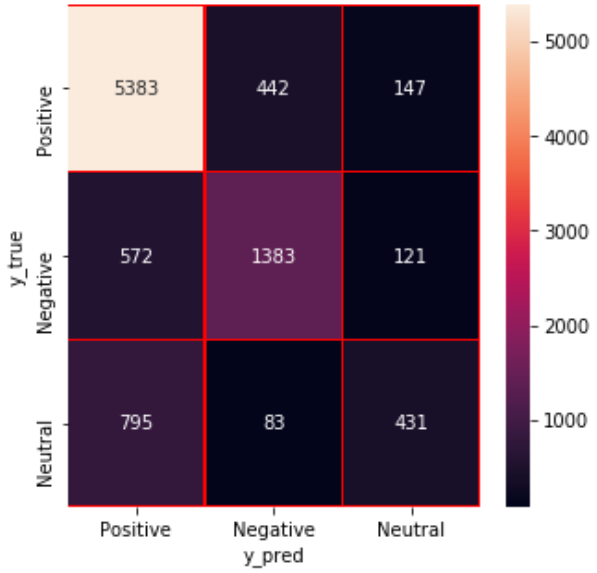
1) 75%의 데이터 (28070개)를 training set으로 사용한다. 나머지 데이터는 test set (9357개) 으로 사용한다.
2) 우리가 예측하고자 하는 것은 감성 라벨이므로, Sentiment(감성 : Positive: 0, Negative: 1, Neutral: 2)를 종속변수로 예측하고자 한다.
3) 가장 많이 쓰이는 단어 3000개에 대하여 감성분석을 진행하였다.

3.2.1 모델링 결과

감성 분석 모델링은 총 5 가지로, Naïve Bayes, RandomForest, SVM, Logistic Regression, Decision Tree 으로 진행하였다.

먼저, Naïve Bayes 모델링의 accuracy 는 76.92%로 측정되었다. 그리고 Naïve Bayes 의 Confusion Matrix 와 Classification Report 는 다음과 같다.

Precision 과 Recall, F1-score 모두 0.75 근방으로 1 에 가까우나 중립 감성(2)에 대해서 0.33 의 Recall 값과 0.43 의 F1-score 값을 보였기 때문에 중립 감성에 대해서는 분류가 정확하지 않은 것으로 보였다. 따라서, 분류가 잘되었다고 보기 어려웠다.

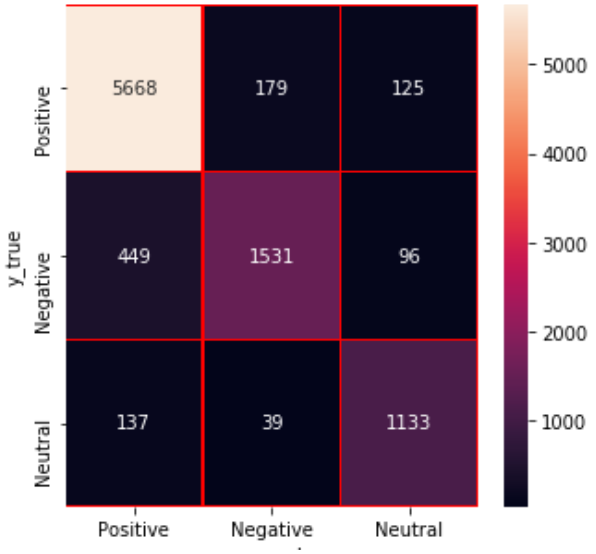


	precision	recall	f1-score	support
0	0.80	0.90	0.85	5972
1	0.72	0.67	0.69	2076
2	0.62	0.33	0.43	1309
avg / total	0.76	0.77	0.75	9357

그림 7: Naïve Bayes Confusion Matrix & Classification Report

RandomForest 모델링의 accuracy 는 89.31%로 측정되었다. 그리고 RandomForest 의 Confusion Matrix 와 Classification Report 는 다음과 같다.

Precision 과 Recall, F1-score 모두 0.9 근방으로 1 에 가깝고, 모든 감성에 대해 0.8 이상의 값을 보였다. 따라서, 분류가 다소 잘되었다고 판단하였다.

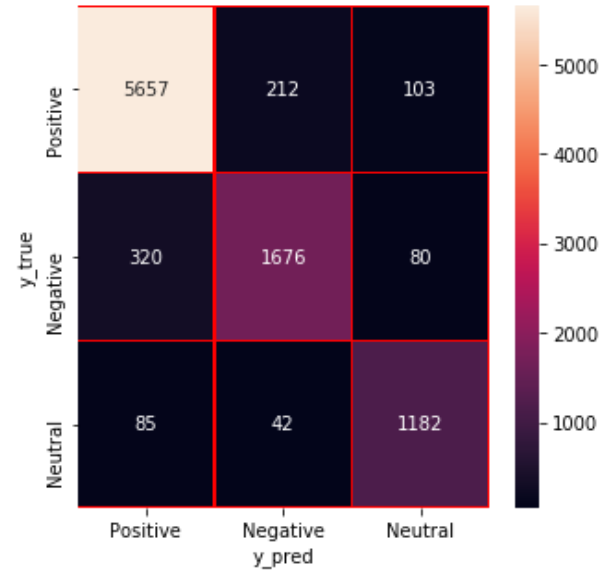


	precision	recall	f1-score	support
0	0.91	0.95	0.93	5972
1	0.88	0.74	0.81	2076
2	0.84	0.85	0.85	1309
avg / total	0.89	0.89	0.89	9357

그림 8: RandomForest Confusion Matrix & Classification Report

SVM 모델링의 accuracy 는 91.00%로 측정되었다. 그리고 SVM 의 Confusion Matrix 와 Classification Report 는 다음과 같다.

Precision 과 Recall, F1-score 모두 0.9 근방으로 1 에 가깝고, 모든 감성에 대해 0.9 이상의 값을 보였다. 따라서, 분류가 다소 잘되었다고 판단하였다.

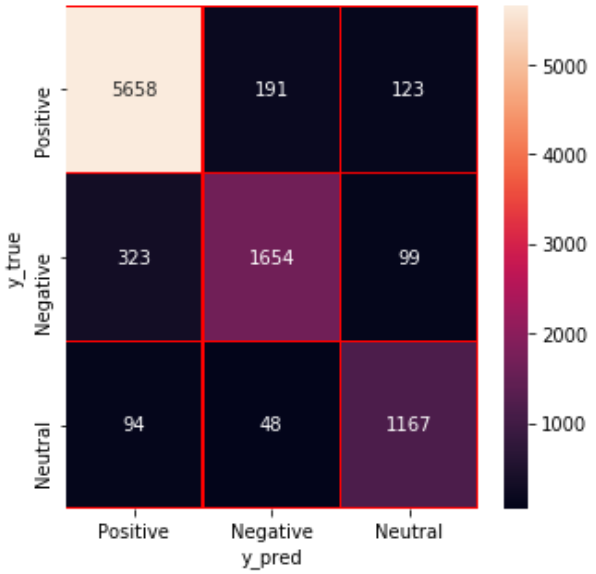


	precision	recall	f1-score	support
0	0.93	0.95	0.94	5972
1	0.87	0.81	0.84	2076
2	0.87	0.90	0.88	1309
avg / total	0.91	0.91	0.91	9357

그림 9: SVM Confusion Matrix & Classification Report

Logistic Regression 모델링의 accuracy 는 90.62%로 측정되었다. 그리고 Logistic Regression 의 Confusion Matrix 와 Classification Report 는 다음과 같다.

Precision 과 Recall, F1-score 모두 0.91 근방으로 1 에 가깝고, 모든 감성에 대해 0.85 근방으로 값이 높다고 판단하였다. 따라서, 분류가 잘되었다고 판단하였다.

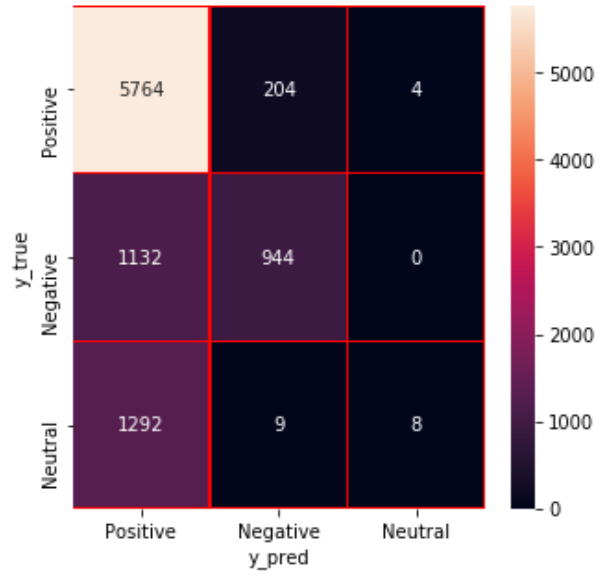


	precision	recall	f1-score	support
0	0.93	0.95	0.94	5972
1	0.87	0.80	0.83	2076
2	0.84	0.89	0.87	1309
avg / total	0.91	0.91	0.91	9357

그림 10: Logistic Regression Confusion Matrix & Classification Report

Decision Tree 모델링의 accuracy 는 71.81%로 측정되었다. 그리고 Decision Tree 의 Confusion Matrix 와 Classification Report 는 다음과 같다.

Precision 과 Recall, F1-score 모두 0.7 근방으로 1 에 가까우나 중립 감성(2)에 대해서 0.01 의 Recall 값과 0.01 의 F1-score 값을 보였기 때문에 중립 감성에 대해서는 분류가 매우 정확하지 않은 것으로 보였다. 따라서, 분류가 잘되지 못했다고 판단하였다.



	precision	recall	f1-score	support
0	0.70	0.97	0.81	5972
1	0.82	0.45	0.58	2076
2	0.80	0.01	0.01	1309
avg / total	0.74	0.72	0.65	9357

그림 11: Decision Tree Confusion Matrix & Classification Report

결과적으로 SVM 모델이 91%로 가장 정확도가 높았다. 그 다음으로는 Logistic Regression 이 90.62%로 높았고, SVM 모델과 큰 차이는 없었다. 따라서, 이 두 가지 모델이 감성 분석 분류 모델링으로 적합하다고 판단하였다.

표 1 : 모델링 정확도 결과 비교표

	Naïve Bayes	RandomForest	SVM	Logistic Regression	Decision Tree
Test Accuracy (%)	76.92	89.31	91.00	90.62	71.81

*SVM 모델 결과 *

가장 정확도가 높았던 SVM 모델로 예측한 결과를 바탕으로 각 감성 별 워드 클라우드를 그려 보았다. 먼저, Positive 리뷰에 대해서는 like, best, great, good, love 등 긍정 감성에 매우 적합한 단어들이 상위에 배치되었다.

반면, Negative 리뷰에 대해서도 like 가 가장 크게 배치되었고, 그 다음은 get, please, help, app work 등이 있었다. like 가 가장 상위 단어인 이유는 아마도 부정문과 같이 싫다는 의미의 단어로 쓰였기 때문이라고 판단하였다. 또한, help, please 와 같이 부정문에 자주 쓰이는 단어들이 출력되어 적절하게 리뷰 데이터가 분류되었다고 판단하였다.

또한, Neutral 리뷰에 대해서 like 가 가장 크게 배치되었으며, 그 다음은 work, get, app 등으로 나타났다.

대부분의 감성을 잘 분류한 것으로 나타났고, 따라서 SVM 모델의 높은 정확도를 살펴볼 수 있었다.

그림 12: SVM 모델에 따른 Positive, Negative, Neutral WordCloud

3.3 LDA 토픽분석

3.3.1 모델링 과정

1) 전체 데이터셋의 리뷰 데이터 모두를 이용하여 LDA 토픽 분석을 진행한다.
2) 데이터 전처리 과정은 앞서 감성분석에서 하였던 전처리 과정과 동일하다.
3) Term-document matrix 생성은 TF-IDF 기준으로 하였다.
4) 토픽 개수는 먼저 30개로 설정한 후, 분포를 보고 줄여나가는 방식으로 진행하였다.

3.3.2 모델링 결과

토픽 모델링 결과 그림은 다음과 같다. 모든 리뷰 데이터에 대하여 모델링한 결과, 30 개의 토픽 중에서도 6~8 개 정도의 군집이 어느 정도 보이는 것을 확인할 수 있었다.

또한, TOP 단어들은 good, nive, love, awesome 등 대부분 긍정 리뷰에 대한 Topic 이었으며, 이것은 긍정 리뷰 데이터의 수가 가장 많기 때문에 당연한 결과라 생각되었다.

어플 카테고리에 해당하는 단어로는, hotel, room, camera, photo, booking 등이 가장 많이 발견되었고, 이것은 숙박 어플이나 카메라 어플 등이 가장 인기가 많은 어플들이라는 것을 나타낸다.

따라서, 전체 리뷰 데이터에 대하여 토픽 개수를 10 개로 설정하고 긍정, 부정, 중립 리뷰 데이터를 각각 나누어 모델링을 다시 진행하였다.

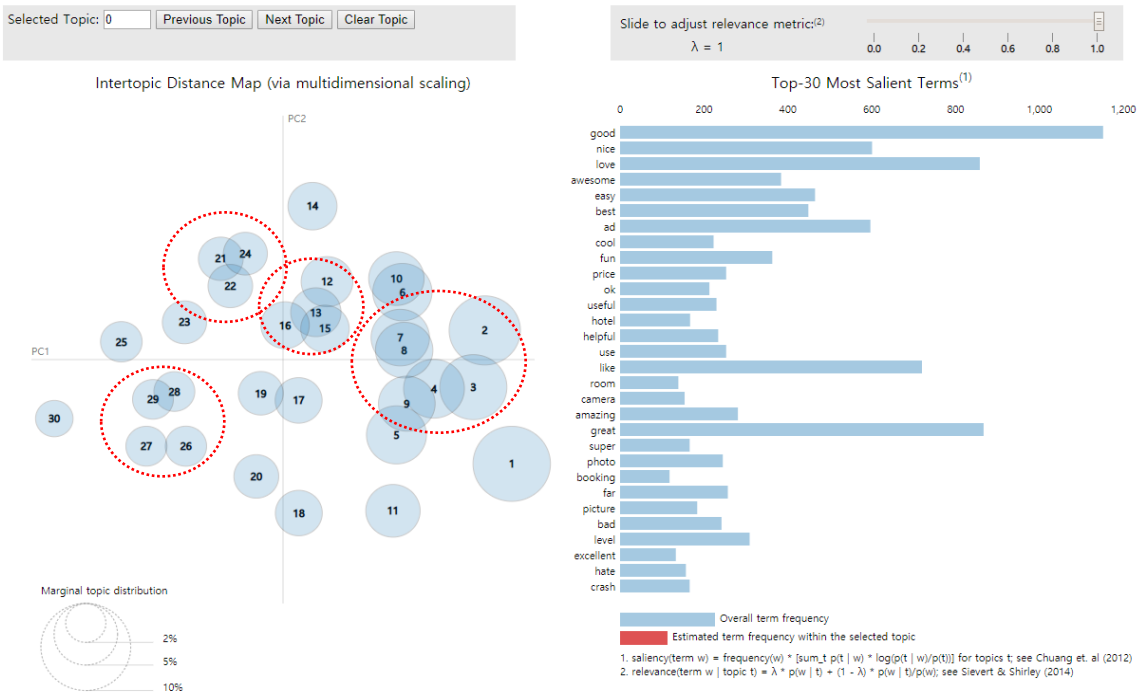


그림 13: 토픽 30 개에 대한 LDA 토픽 모델링 결과

토픽 10 개에 대한 긍정 리뷰 데이터 토픽 모델링 결과 그림은 다음과 같다.

또한, TOP 단어들은 good, great, useful, amazing 등 대부분 좋은 반응 Topic 단어였다.

어플 카테고리에 해당하는 단어로는, game, business, camera, profile 등이 가장 많이 발견되었고, 이것은 게임 어플이나 비즈니스, 카메라 어플 등이 가장 인기가 많은 어플들이라는 것을 나타낸다.

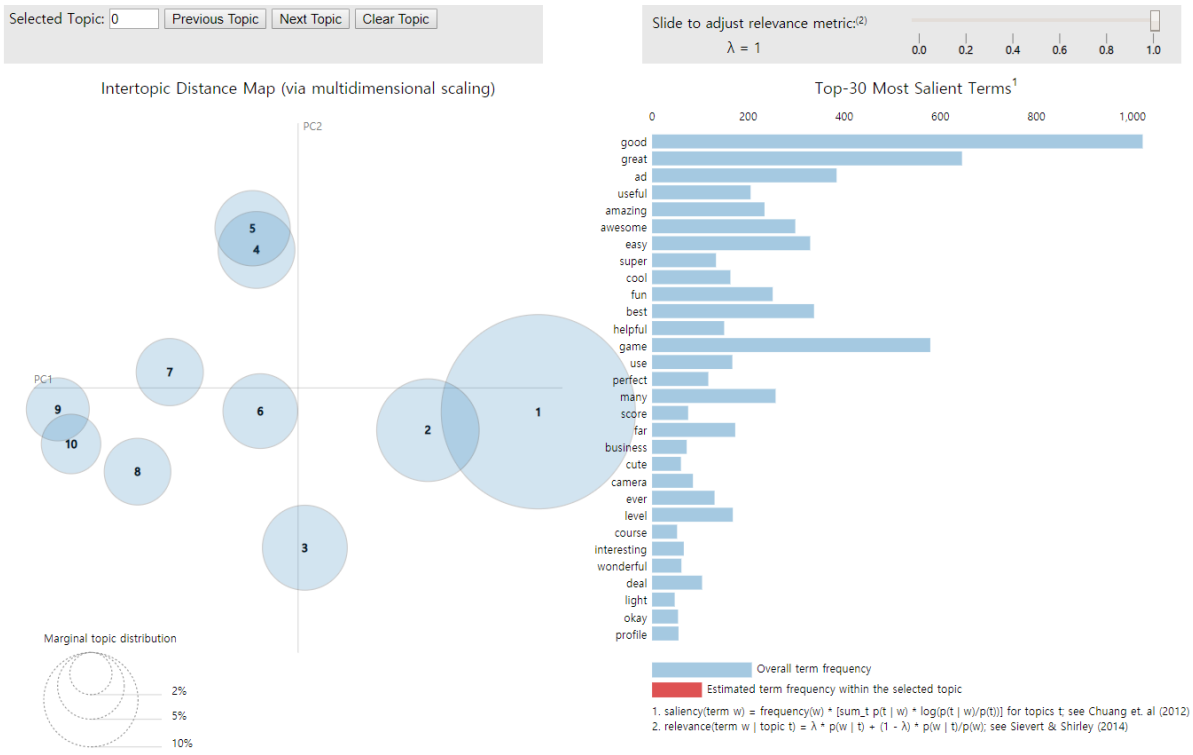


그림 14: 토픽 10 개에 대한 긍정리뷰 LDA 토픽 모델링 결과

토픽 10 개에 대한 부정 리뷰 데이터 토픽 모델링 결과 그림은 다음과 같다.

또한, TOP 단어들은 hotel, agent, slow, worst, room, booking 등이었다. 이것들은 모두 숙박 어플에 대한 단어들로, 대부분의 부정 리뷰들이 숙박과 관련된 것들이라는 것을 알 수 있었다. 따라서, 다른 어플들보다도 숙박 관련 어플들은 리뷰에 대한 대응을 신속하게 할 필요가 있다고 생각하였다.

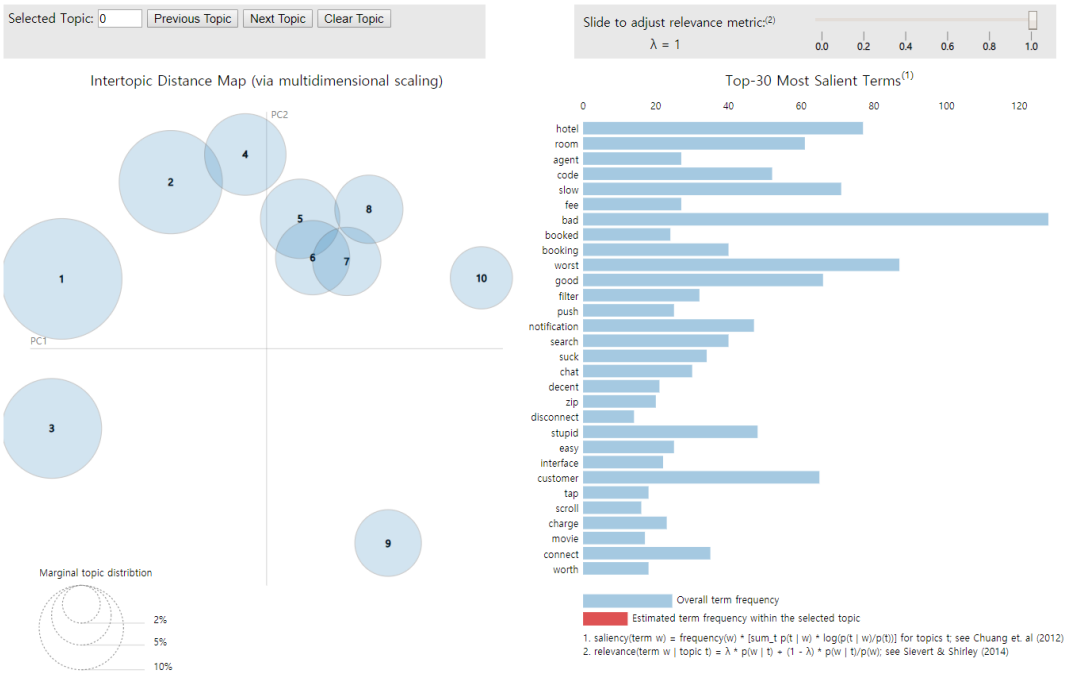


그림 15: 토픽 10 개에 대한 부정리뷰 LDA 토픽 모델링 결과

토픽 10 개에 대한 중립 리뷰 데이터 토픽 모델링 결과 그림은 다음과 같다.

또한, TOP 단어들은 like, thank, work, helpful, nyc 등이고, 별다른 특징은 발견할 수 없었다.

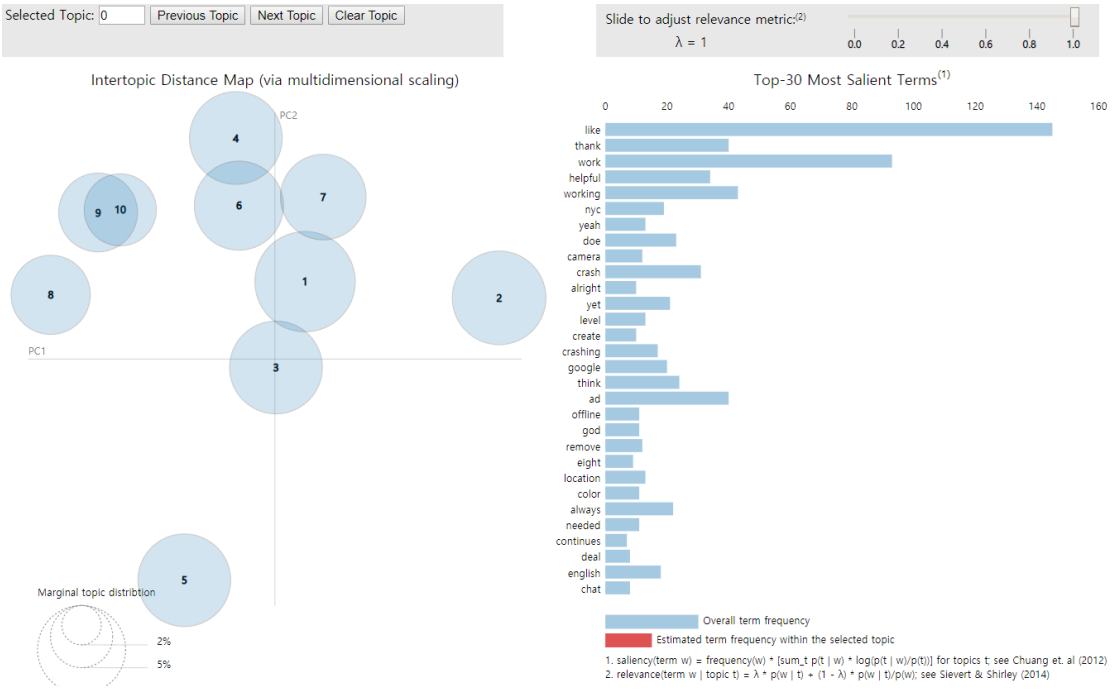


그림 16 : 토픽 10 개에 대한 중립리뷰 LDA 토픽 모델링 결과

4. 결과 및 개선 방안

4.1 기대효과

구글 앱스토어 리뷰 데이터를 활용하여 감성분석 및 LDA 토픽분석을 진행하였다. 감성분석 모델링에서는 SVM 모델이 가장 성능이 좋았고, 그 다음은 Logistic Regression 모델이었다. 이러한 감성분석 모델링을 통해, 구글 앱스토어 리뷰 데이터에 대해서는 어떤 감성이 가장 많이 분포하며, 분류 모델로는 무엇이 적합한지 알아볼 수 있었다.

LDA 토픽 모델링에서는 리뷰 데이터에 내재된 토픽들을 추출하여 감추어진 화제 혹은 정리된 개념을 산출할 수 있었다. 또한, LDA 토픽 모델을 통해 특정 토픽에 특정 단어가 나타날 확률을 제시하거나 각 리뷰 데이터의 토픽 분포를 제시할 수 있었다. 이러한 텍스트 데이터의 토픽을 추출함을 통해 해당 어플 사용자의 관심사나 여론을 분석할 수 있을 것이다.

4.2 개선 방안

구글 앱스토어 리뷰 데이터의 감성분석에서 Neural Network 모델을 도입한다면, SVM 모델 보다 좀 더 높은 성능을 보일 것이라 예상한다. 또한, 다른 모델들에 대해서도 Tuning parameter 를 달리 한다면, 좀 더 나은 정확도를 보일 것이라 생각한다.