

Data Analytics

Individual Assignment 2

202011431 산업공학과 차승현

배달 어플리케이션 리뷰 데이터에 감성분석 기법을
적용한 소비자 만족도 제고 방향성 제안



목차

I. 서론	3
II. 분석 기법	5
III. 데이터 전처리 및 분석과정	8
IV. 분석 결과 해석	34
V. 결론 및 한계점	41
References	44

I. 서론

오늘날 SNS(Social Network Service)와 전자상 거래, 온라인 커뮤니티의 발전으로, 온택트 소비과정이 증가함에 따라 소비자들은 자신이 구매했던 상품과 서비스에 대한 평가를 적극적으로 작성하게 되었고 이렇게 축적된 상품 리뷰 데이터는 다른 소비자들의 구매 행위에 매우 큰 영향을 미치고 있다. 소비자들은 상품을 구매하기 전 해당 상품의 리뷰 페이지에 접속하여 기존 사용자들의 평가를 살펴본 후 자신의 구매의사를 결정한다. 뿐만 아니라 다른 고객의 실사용 후기가 더 객관적이며, 신뢰할 만하다고 인지한다. 따라서 리뷰는 상품 구매의 트리거 역할을 할 뿐만 아니라 마케팅 데이터 혹은 상품 개선을 위한 지표가 되므로 매우 가치 있는 정보이다.

이러한 O2O(Online to Offline) 서비스 기반의 배달 애플리케이션이 성장할 수 있었던 것은 사회적, 환경적 변화와 기술의 발전으로 설명할 수 있다(김민선, 2020). 첫 번째, 사회적으로 1~2인 중심의 가구 재편이 이루어지고, 맞벌이 부부 증가 및 외부 활동 시간 증가로 인해 간편하게 식사를 해결하려는 소비자가 늘어나 배달 외식 수요가 증가한 것으로 나타났다(한국농수산식품유통공사, 2015). 두 번째, 환경적으로 미세먼지나 폭염과 같이 외부 활동에 제약을 주는 요소로 집에서 간편하게 식사를 해결하려는 수요가 배달 외식을 증가시킨 것으로 나타났다(오픈서베이, 2019). 마지막으로 정보통신기술(ICT) 발전과 스마트폰 보급의 확산을 바탕으로 O2O 서비스 기반의 배달 애플리케이션이 등장한 것이다. 새롭게 등장한 배달 애플리케이션으로 인해 소비자의 스마트폰 위치 정보를 활용하여 실시간으로 배달이 가능한 주변 음식점과 매장 정보, 메뉴 정보, 소비자 후기, 간편 결제 시스템 등 간편한 서비스를 소비자에게 제공할 수 있게 되었다. 배달 애플리케이션은 사회적, 환경적 변화로 증가한 배달 외식의 수요와 간편하게 식사를 해결하려는 소비자들의 니즈를 충족시키며 큰 호응을 얻게 된 것이다.

2019년 12월 말 발생한 코로나19도 배달 외식 수요 증가와 배달 애플리케이션 활성화에 큰 영향을 미쳤다. 질병 확산 방지를 위해 조치한 사회적 거리 두기 운동의 여파로 집콕족이 늘어나고, 비대면 소비문화가 확산되면서 O2O 기반의 배달

애플리케이션 시장의 수요가 급격히 증가한 것이다.

이처럼 배달 어플리케이션 시장의 수요가 증가함에 따라서 배달 시장의 공급도 자연스럽게 증가하며, 배달 어플리케이션 시장의 규모는 점차 커지고 있다. 따라서 배달 어플리케이션끼리의 경쟁도 심화되고 있으며, 한 배달 어플리케이션 내에서의 가게들끼리의 경쟁들도 치열해지고 있는 현실이다. 방대한 배달 요청의 양에 가게와 라이더(배달 기사원)가 신경쓰지 못하는 사소한 부분에서 소비자들은 여러 불만을 느끼곤 한다. 배달 애플리케이션 시장 성장 및 경쟁과 더불어 보완이 필요한 여러 관점에서의 다양 한 문제점도 드러나고 있다. 예를 들어, 환경적으로는 음식 포장을 위해 사용하는 일회용품과 일회용품 처리에 대한 문제가 있다. 이를 두고 배달의 민족, 요기요, 쿠팡 이츠는 일회용 식기 사용 줄이기 캠페인을 공동으로 진행한다고 밝히기도 하였다(연합뉴스, 2021). 소상공인 관점으로는 배달 애플리케이션 기업에 지급하는 광고료나 수수료 문제, 갑질 등의 문제가 있다. 이에 공공 지자체에서 출시하는 지역 배달 애플리케이션의 경우 비교적 낮은 수수료를 부과하고 있으며, 배달의 민족은 한때 일방 적으로 발표했던 수수료 정책을 철회하기도 하였다. 소비자 관점으로는 배달 애플리케이션 기업마다 민원 유형이 조금씩 다르지만, 음식 미배달이나 오배달, 주문 취소, 품질 등 다양하게 지적되고 있다. 2018년 소비자 민원 평가 집계된 자료에 따르면 시스템 오류로 발생한 소비자 민원이 36.8%를 차지하는 것으로 나타났다(소비자가 만드는 신문, 2020). 주문 누락 및 취소를 의미하는 시스템 오류는 소비자가 주문 후 수습 분을 기다린 후에 사실을 인지하다 보니 소비자 불만이 증폭되는 경향을 보였다. 시스템 오류 이외에도 취소, 환불 및 서비스에 대한 민원이 접수되는 것으로 나타났다. 이처럼 여러 관점에서 다양한 불만이 제기되는 것은 배달 애플리케이션의 수요가 늘어나고 시장이 급격히 성장하는 과정에서 해결되지 않은 문제들이 드러난 것으로 판단되며, 해결방안 모색이 필요한 시점으로 보인다.

따라서 본 보고에서는 배달 어플리케이션 ‘요기요’ 사용자 리뷰를 분석하고 소비자 만족 및 불만족에 영향을 미치는 요인 및 키워드를 파악하여 소비자들의 니즈를 충족시키고 동시에 시장 경쟁력 제고할 수 있는 방향성을 제언하고자 한다.

II. 분석 기법

1) 감성 분석(Sentiment analysis)

감성분석은 텍스트 마이닝 기법 중 하나로 텍스트 문서에 포함되는 다양한 극성과 감성을 추출하는 방식이다. 일반적으로 텍스트에 대한 긍정 혹은 부정적 언어를 식별하고 분류하는 텍스트 분류 문제에 많이 활용되고 있다(Hu et al., 2012). 감성 분석 방법은 어휘기반, 기계 학습 및 하이브리드 방법의 세 가지 유형으로 분류할 수 있다(Ravi & Ravi, 2015). 어휘기반 감성분석은 감성사전에 의해 이루어지며, 각 문서의 단어를 감성사전의 어휘와 매칭한다. 기계학습 방법은 텍스트에 수동으로 레이블을 지정하는 과정이 필요하며 모델과 데이터에 크게 의존하기 때문에 어휘기반 감성분석이 보다 효율적인 방법이라 할 수 있다(Zhu et al., 2020). 감성분석은 텍스트로 표현된 의견과 감성을 식별할 수 있고 온라인 리뷰를 통한 소비자의 긍정이나 부정적인 제품 및 서비스에 대한 평가를 구분할 수 있다. Sharma et al.(2020)은 감성분석을 통해 리뷰 평점과 리뷰 감성 간의 관계를 분석하였으며 소비자가 제공받은 호텔의 서비스와 기대한 서비스의 차이가 감성에 영향을 미친다는 것을 밝혔다. Tsai et al.(2020)은 호텔 리뷰 중 유용한 정보를 정확하게 추출하기 위하여 감성분석을 기반으로 자동 리뷰 요약 시스템을 구축하였다.

한국어 감성분석은 딥러닝 기법이 점점 발달하면서 딥러닝을 활용한 연구가 주로 이루어지고 있다. 기존의 기계학습 방식은 학습데이터의 특성에 영향을 많이 받아 도메인 적응에 취약해 저조한 성능을 보이는데 이를 해결하기 위해 딥러닝 기법을 적용하면 학습데이터에서 높은 수준으로 데이터를 추출하여 뛰어난 성과를 얻을 수 있다.

2) TF-IDF(빈도 가중치 모델)

TF-IDF(Term Frequency - Inverse Document Frequency)는 문서 내에서 특정 단어가 갖는 중요도를 나타내는 가중치이다. 문서 내에서 특정 단어가 출현하는 빈도와 특정 단어를 갖고 있는 문서가 전체 문서에서 차지하는 비율의 역수에 로그를 취한 값을 곱하여 값을 구한다. 단어의 출현 빈도만으로 단어의 가중치를 결정할

때 갖는 한계를 문서의 역빈도를 통해 개선한 방법이다. TF-IDF는 텍스트 마이닝 연구에 많이 쓰이는 방법 중에 한가지로 리뷰, 기사, 연설문 등을 대상으로 감정분석, 검색, 문서 분류, Keyword 추출과 관련한 연구가 활발히 진행 중이다. 최근 SNS, 온라인 커뮤니티, 미디어의 발달로 개인이 데이터를 생산하는 주체가 되었다. 이러한 변화로 인해 텍스트를 비롯한 비정형데이터의 잠재력과 중요성이 커지고 있다. 최근에는 인공지능망을 이용한 자연어처리 등 많은 연구가 진행되고 있다.

3) Logistic Regression

로지스틱 회귀분석법은 Cox 에 의해 제안된 확률모델이며 종속변수가 이항형일 경우 설명변수의 선형결합을 이용하여 사건의 발생 가능성을 예측하는 데 널리 사용되는 통계기법으로 자료 분류에 그 목적이 있다. 로지스틱 회귀분석은 각 변수의 영향력을 알 수 있다는 장점이 있지만 한 설명변수의 효과가 다른 설명변수의 수준에 의존하지 않는다고 가정하기 때문에 예측 모델에 투입된 여러 설명변수들 간 상호작용의 모든 경우의 수를 고려하여 분석하는 것은 거의 불가능하다.

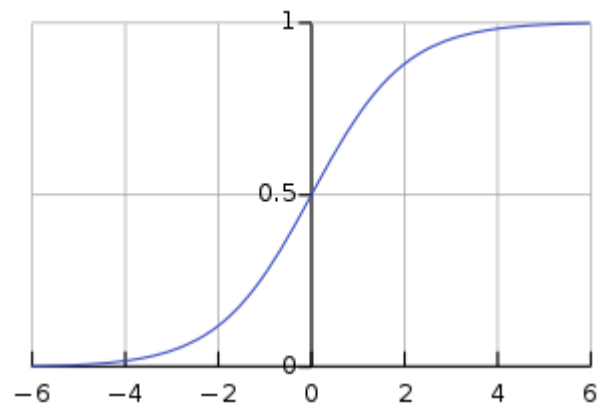
로지스틱 회귀는 이항형 또는 다항형이 될 수 있다. 이항형 로지스틱 회귀(binomial logistic regression)의 경우 종속 변수의 결과가 (성공, 실패) 와 같이 2 개의 카테고리가 존재하는 것을 의미하며 다항형 로지스틱 회귀는 종속형 변수가 (맑음, 흐림, 비)와 같이 2 개 이상의 카테고리로 분류되는 것을 가리킨다. 이항형 로지스틱의 회귀 분석에서 2 개의 카테고리는 0 과 1 로 나타내어지고 각각의 카테고리로 분류될 확률의 합은 1 이 된다.

로지스틱 회귀는 일반적인 선형 모델(generalized linear model)의 특수한 경우로 볼 수 있으므로 선형 회귀와 유사하다. 하지만, 로지스틱 회귀의 모델은 종속 변수와 독립 변수 사이의 관계에 있어서 선형 모델과 차이점을 지니고 있다. 첫 번째 차이점은 이항형인 데이터에 적용하였을 때 종속 변수 y 의 결과가 범위[0,1]로 제한된다는 것이고 두 번째 차이점은 종속 변수가 이진적이기 때문에 조건부 확률($P(y | x)$)의 분포가 정규분포 대신 이항 분포를 따른다는 점이다.

따라서, 대상이 되는 데이터의 종속 변수 y 의 결과는 0과 1, 두 개의 경우만 존재하는 데 반해, 단순 선형 회귀를 적용하면 범위[0,1]를 벗어나는 결과가 나오기 때문에 오히려 예측의 정확도만 떨어뜨리게 된다.

이를 해결하기 위해 로지스틱 회귀는 연속이고 증가함수이며 [0,1]에서 값을 갖는 연결 함수 $g(x)$ 를 제안하였다. 연결함수의 형태는 다양하게 존재하는데 그 중 대표적인 두 개는 아래와 같다.

- 로지스틱 모형: $g(x) = \frac{e^x}{1 + e^x}$



III. 데이터 전처리 및 분석과정

분석에 사용할 데이터는 다음과 같다.

구분	설명
모집단	배달 어플리케이션 ‘요기요’ 온라인 리뷰를 작성한 사용자
표본 집단	배달 어플리케이션 ‘요기요’의 화양동에 위치한 매장의 리뷰 데이터
표본 크기	배달 어플리케이션 ‘요기요’의 리뷰 28,917개
샘플링 조건	2018.05.09 이전에 작성된 리뷰수가 500건 이상인 13개의 업체에 대한 리뷰

Python을 사용하여 제공된 Data를 불러왔다. Data의 형태는 다음과 같다.

	업체명	카테고리	메뉴	맛	양	배달	리뷰	date
0	전주석식물고기-본점	한식	파절이매콤통삼겹 (2~3인) (공기밥2+김치찌개+밀반찬+쌈) /1	5.0	5.0	5.0	자주시켜먹는 단골집인데 항상변치않고 맛있습니다!!	2017년 12월 6일 수요일
1	전주석식물고기-본점	한식	통삼겹살 2인 (고기+공기밥2+김치찌개+쌈+밀반찬) /1(추가 선택(고기 추가))	5.0	5.0	5.0	배달 시간도 오래걸리지 않고, 양이 적을거 같아서 고기 추가를 했는데..안해도 됐었...	2017년 9월 30일 토요일
2	전주석식물고기-본점	한식	통삼겹살 2인 (고기+공기밥2+김치찌개+쌈+밀반찬) /1	NaN	NaN	NaN	굿굿	2017년 9월 23일 토요일
3	전주석식물고기-본점	한식	통삼겹 (小 / 500g) (냉얼무국수 or 냉얼무우동 or 비빔얼무국수+쌈) /1(메뉴 선택...)	5.0	5.0	5.0	배달도빠르고맛나요	2018년 3월 13일 화요일
4	전주석식물고기-본점	한식	통삼겹 (3~4인) (김치찌개+공기밥3+밀반찬+쌈) /1	5.0	5.0	5.0	찌개와 통삼겹 맛있는 5잔. 많이 먹는편이라 3인분짜리 주문했는데...배터지는줄...	2018년 2월 27일 화요일
...
28912	동강	중식		NaN	NaN	NaN	좋아요	2013년 12월 31일 화요일
28913	동강	중식		NaN	NaN	NaN	맛있ㅇㅜㅌ	2013년 12월 29일 일요일
28914	동강	중식		NaN	NaN	NaN	맛이게다	2013년 12월 25일 수요일
28915	동강	중식		NaN	NaN	NaN	여기 맛있어요 이근저 향수옥은 냄새?ㅜㅜ돼지냄새? 나는 곳이 너무 많은데 여긴안그래요...	2013년 12월 10일 화요일
28916	동강	중식		NaN	NaN	NaN	맛있어요 또시켜먹어야지 추천할게요	2013년 11월 13일 수요일

3-1) Raw Data의 기초 통계량 및 결측치 확인

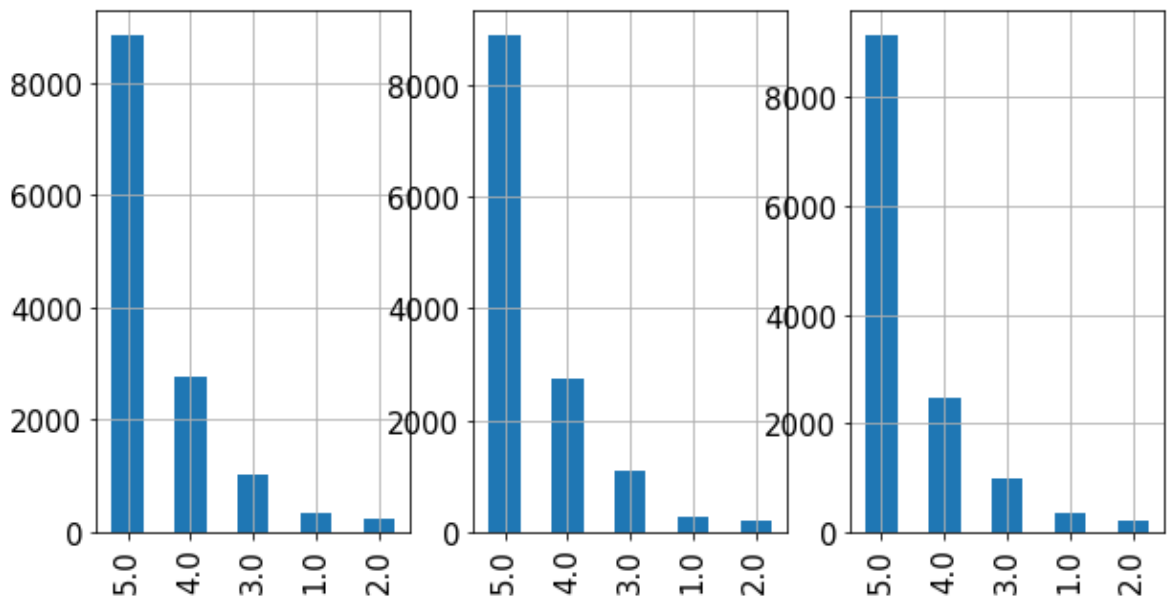
	맛	양	배달
Count	13196	13196	13196
Mean	4.479388	4.497196	4.498636
Std	0.903926	0.867802	0.916067
Min	1.0	1.0	1.0
Q1	4.0	4.0	4.0
Q2	5.0	5.0	5.0
Q3	5.0	5.0	5.0
Max	5.0	5.0	5.0

Raw Data의 기초 통계량을 관측하였을 때 리뷰 텍스트 데이터의 기반이 될 수 있는 세부 평가항목(맛, 양, 배달)의 평가 점수가 매우 유사함을 확인할 수 있었다. 또한, Data의 전체 개수는 28917개이지만 분석의 기반이 되는 ‘맛’, ‘양’, ‘배달’, ‘리뷰’ data가 NaN인 경우가 존재했다.

	맛	양	배달	리뷰
NaN	15721	15721	15721	7

Raiting	맛	양	배달
1	331 (0.02508%)	260 (0.01970%)	367 (0.02781%)
2	241 (0.01826%)	222 (0.01682%)	229 (0.01735%)
3	1035 (0.07843%)	1093 (0.08283%)	998 (0.07563%)
4	2753 (0.2086%)	2743 (0.20787%)	2465 (0.18680%)
5	8836 (0.6696%)	8878 (0.67278%)	9137 (0.69241%)

추가로, 맛과 양 배달의 점수 부여 기준이 매우 유사함을 알 수 있었다.



맛-양-배달 항목의 점수 분포도 그래프

결측치를 제거하고 데이터를 분석하는 것이 제일 좋은 성능을 보이겠지만, ‘맛’, ‘양’, ‘배달’ 항목의 결측치가 동등한 것으로 보아 셋 중 하나의 항목에 결측치가 하나라도 존재한다면 나머지 두 항목의 평가도 결측치인 것을 확인할 수 있었다. 따라서 결측치를 어느 특정한 값으로 대체해야 한다고 판단하였다. 맛, 양, 배달 항목의 결측치는 리뷰 항목의 텍스트 데이터에서 긍정 단어 및 부정단어를 추출하여 예측할 수 있지만, 리뷰 데이터의 결측치는 쉽게 예측이 불가능하며, 데이터상의 오류를 초래할 수 있다. 따라서 리뷰 데이터의 결측치가 존재하는 7개의 데이터는 삭제하고 데이터를 분석하기로 결정했다.

	업체명	카테고리		메뉴	맛	양	배달	리뷰	date
694	피자샵-자양성수점	피자		하프엔하프피자 L/1(피자 선택(수제고구마,통마늘불고기),도우 선택(리치글드)),수...	5.0	4.0	4.0	NaN	2018년 2월 11일 일요일
772	피자샵-자양성수점	피자		페파로니피자 XXXL/1(도우 선택(치즈크러스트)),코카콜라 1.25L/1	4.0	5.0	5.0	NaN	2018년 2월 2일 금요일
802	피자샵-자양성수점	피자		콤비네이션피자 XXXL/1(도우 선택(오리지널))	5.0	5.0	5.0	NaN	2018년 1월 28일 일요일
939	피자샵-자양성수점	피자		수제고구마피자 L/1(도우 선택(소보로)),코카콜라 1.25L/1	5.0	4.0	1.0	NaN	2018년 2월 27일 화요일
1022	피자샵-자양성수점	피자		세트2 (오리지널피자L + 사이드1 + 사이드2 + 콜라1.25L) /1(피자 선택(쉬림프),도우...	5.0	5.0	1.0	NaN	2018년 1월 21일 일요일
1157	피자샵-자양성수점	피자		불고기피자 L/1(도우 선택(치즈크러스트)),치킨텐더 (4조각) /1,체다웨이감자/1,...	5.0	5.0	4.0	NaN	2018년 1월 31일 수요일
1235	피자샵-자양성수점	피자		NaN	5.0	5.0	3.0	NaN	2018년 2월 24일 토요일

3-2) 데이터 전처리

데이터 전처리의 방향성은 다음과 같다.

맛, 양, 배달의 평가항목의 결측치가 존재하고, 동시에 리뷰 항목에 특정 단어가 포함되어 있다면 해당 단어가 들어가는 리뷰들의 맛, 양, 배달의 산술평균을 통해서 소비자의 전체적인 만족감, 즉 Total score를 매기고자 한다. 예를 들어, ‘너무너무 맛나요’ 라는 데이터의 맛, 양, 배달의 평가항목이 결측치라면 ‘맛나요’ 라는 단어가 들어가는 사건과, 맛, 양, 배달의 평가항목이 결측치가 아닌 사건의 교집합의 산술평균으로 결측치를 대체하고자 한다. 그러나 Raw Data의 리뷰를 살펴보면 ‘맛 있어용!!’, ‘맛이써용ㅎ’, ‘넘나맛나는것’ 등 ‘맛있다’를 표현하기 위해서 사전에 정의된 단어가 아닌 띄어쓰기 및 문법과 대소문자 구분이 무시되거나, 특수문자 및 기호가 포함되어 작성되어 있는 경우가 많았다. 따라서 리뷰 데이터의 띄어쓰기를 전부 제거하여 ‘맛 있다’와 ‘맛있다’를 동일한 의미로 해석이 되게끔 정제한다. 또 개별 data의 row마다 고유 key를 부여하여, 리뷰 항목의 텍스트 데이터에 띄어쓰기를 제거하고 맛, 양, 배달 항목이 결측치이지만 리뷰 데이터의 특정 단어를 포함하여 Total score를 산출하는 정제과정을 거친 후의 data를 raw data와 결합하여 기존의 리뷰데이터에 작성된 정제과정을 거치지 않은(띄어쓰기가 존재하는) 리뷰 텍스트 데이터와 total score를 결합하는 방식을 고안하였다.

	업체명	카테고리	메뉴	맛	양	배달	리뷰	date	numbering
0	전주석식물고기-본점	한식	파절이매콤통삼겹 (2~3인) (공기밥2 + 김치찌개 + 밑반찬 + 찜) /1	5.0	5.0	5.0	자주시켜먹는 단골집인데 항상변치않고 맛있습니다!!	2017년 12월 6일 수요일	0
1	전주석식물고기-본점	한식	통삼겹살 2인 (고기 + 공기밥2 + 김치찌개 + 찜 + 밑반찬) /1(추가 선택(고기 추가))	5.0	5.0	5.0	배달 시간도 오래걸리지 않고, 양이 적을거 같아서 고기 추가를 했는데..안해도 됐었...	2017년 9월 30일 토요일	1
2	전주석식물고기-본점	한식	통삼겹살 2인 (고기 + 공기밥2 + 김치찌개 + 찜 + 밑반찬) /1	NaN	NaN	NaN	굿굿	2017년 9월 23일 토요일	2
3	전주석식물고기-본점	한식	통삼겹 (小 / 500g) (냉열무국수 or 냉열무우동 or 비빔열무국수 + 찜) /1(메뉴 선택...)	5.0	5.0	5.0	배달도빠르고맛나요	2018년 3월 13일 화요일	3
4	전주석식물고기-본점	한식	통삼겹 (3~4인) (김치찌개 + 공기밥3 + 밑반찬 + 찜) /1	5.0	5.0	5.0	찌개와 통삼겹 맛있는 5찬. 많이 먹는편이라 3인분짜리 주문했는데... 배터지는줄 ...	2018년 2월 27일 화요일	4
...
28912	동강	중식		NaN	NaN	NaN	좋아요	2013년 12월 31일 화요일	28912
28913	동강	중식		NaN	NaN	NaN	맛있ㅇ ㅊㅌ	2013년 12월 29일 일요일	28913
28914	동강	중식		NaN	NaN	NaN	맛이게다	2013년 12월 25일 수요일	28914
28915	동강	중식		NaN	NaN	NaN	여기 맛있어요 이근저 탕수육은 냄새?ㅈㅈ돼지 냄새?나는 곳이 너무 많은데 여긴안그래요...	2013년 12월 10일 화요일	28915
28916	동강	중식		NaN	NaN	NaN	맛있어요 또시켜먹어야지 추천할게요	2013년 11월 13일 수요일	28916

개별 data에 ‘numbering’이라는 column을 생성하여 고유 key를 부여한 모습을 볼 수 있다. 위 data에서 리뷰 항목에 결측치가 있는 행들을 제거하였다.

	업체명	카테고리	메뉴	맛	양	배달	리뷰	date	numbering	total
0	전주석식물고기-본점	한식	파절이매콤통삼겹 (2~3인) (공기밥2 + 김치찌개 + 밑반찬 + 찜) /1	5.0	5.0	5.0	자주시켜먹는 단골집인데 항상변치않고 맛있습니다!!	2017년 12월 6일 수요일	0	5.0
1	전주석식물고기-본점	한식	통삼겹살 2인 (고기 + 공기밥2 + 김치찌개 + 찜 + 밑반찬) /1(추가 선택(고기 추가))	5.0	5.0	5.0	배달 시간도 오래걸리지 않고, 양이 적을거 같아서 고기 추가를 했는데..안해도 됐었...	2017년 9월 30일 토요일	1	5.0
2	전주석식물고기-본점	한식	통삼겹살 2인 (고기 + 공기밥2 + 김치찌개 + 찜 + 밑반찬) /1	NaN	NaN	NaN	굿굿	2017년 9월 23일 토요일	2	NaN
3	전주석식물고기-본점	한식	통삼겹 (小 / 500g) (냉열무국수 or 냉열무우동 or 비빔열무국수 + 찜) /1(메뉴 선택...)	5.0	5.0	5.0	배달도빠르고맛나요	2018년 3월 13일 화요일	3	5.0
4	전주석식물고기-본점	한식	통삼겹 (3~4인) (김치찌개 + 공기밥3 + 밑반찬 + 찜) /1	5.0	5.0	5.0	찌개와 통삼겹 맛있는 5찬. 많이 먹는편이라 3인분짜리 주문했는데... 배터지는줄 ...	2018년 2월 27일 화요일	4	5.0
...
28912	동강	중식		NaN	NaN	NaN	좋아요	2013년 12월 31일 화요일	28912	NaN
28913	동강	중식		NaN	NaN	NaN	맛있ㅇ ㅊㅌ	2013년 12월 29일 일요일	28913	NaN
28914	동강	중식		NaN	NaN	NaN	맛이게다	2013년 12월 25일 수요일	28914	NaN
28915	동강	중식		NaN	NaN	NaN	여기 맛있어요 이근저 탕수육은 냄새?ㅈㅈ돼지 냄새?나는 곳이 너무 많은데 여긴안그래요...	2013년 12월 10일 화요일	28915	NaN
28916	동강	중식		NaN	NaN	NaN	맛있어요 또시켜먹어야지 추천할게요	2013년 11월 13일 수요일	28916	NaN

또 ‘total’이라는 column을 맛, 양, 배달 항목의 산술평균으로 산출하여 생성하였다. 위 세가지 항목들에 결측치가 있다면 Total Score 또한 결측치로 산정됨을 확인할 수 있다. 이처럼 결측치가 존재하는 data에 total score를 산정하기 위해

‘특정 단어’를 포함한 리뷰 데이터라면 그 단어를 포함하고 동시에 결측치가 아닌 data의 total score의 평균값으로 대체하려고 한다.

	‘특정 단어’에 포함되는 단어	해당 단어가 포함되고, 평가가 없는 data의 개수	해당 단어가 포함되고, 평가가 있는 data의 total score의 mean
맛 관련된 직접적인 좋은 평가	‘맛있’, ‘맛잇’, ‘마싯’, ‘맛이’, ‘맛도’, ‘마싯’, ‘마이’, ‘맛남’, ‘맛나’, ‘꿀맛’, ‘맛있’	9402	4.514
‘좋음’을 표현하는 평가	‘좋아’, ‘좋네’, ‘좋다’, ‘좋’, ‘조아’, ‘조와’, ‘굿’, ‘굳’, ‘너무’, ‘엄청’, ‘넘’, ‘존나’, ‘졸라’, ‘존맛’, ‘졸맛’, ‘good’, ‘나이스’, ‘명불허전’, ‘구프’, ‘nice’, ‘ood’, ‘인정’, ‘깔끔’, ‘단골’, ‘딱’	2436	4.5
서비스에 관련된 좋은 평가	‘친절’, ‘훌륭’, ‘추천’, ‘편하’, ‘대박’, ‘짱’, ‘ㅎ’, ‘ㅋ’, ‘최고’, ‘진리’, ‘잘 먹’, ‘잘’, ‘만족’, ‘번창’, ‘감사’, ‘시켜먹으’, ‘시켜드세’, ‘밥도둑’, ‘배불’, ‘배부’, ‘빠른’, ‘빨’, ‘빨리’, ‘빠르고’, ‘빠’, ‘믿고’, ‘믿’, ‘여기만’	1546	4.514
중립을 표현하는 평가	‘그닥’, ‘무난’, ‘평범’, ‘그저그’, ‘쏘쏘’, ‘보통’, ‘괜’, ‘갠’, ‘그럭’, ‘생각보’, ‘별로’, ‘그냥’, ‘먹을만’, ‘평타’	669	4.457
나쁨을 표현하는 평가	‘별로’, ‘맛없’, ‘노맛’, ‘없’, ‘않’, ‘마세’, ‘안’, ‘마라’, ‘최악’, ‘아깝’, ‘실망’, ‘πππ’, ‘않’, ‘마세’, ‘마셈’, ‘마삼’, ‘짜증’, ‘——’, ‘돈아까’, ‘돈아깝’, ‘아까’, ‘똥’, ‘배고’, ‘적다’	517	4.439

음식 및 서비스에 대한 나뭇잎을 표현하는 평가	‘짹’, ‘맴’, ‘짜’, ‘매워’, ‘재탕’, ‘딱딱’, ‘비싸’, ‘누락’	92	4.44
---------------------------	---	----	------

‘특정 단어’에 포함되는 단어의 선별 기준은 value_counts() 함수를 사용하여 대체적으로 많이 사용되는 키워드들을 우선적으로 넣은 후 결측치를 반복하여 도출하였다. 분석을 수행하면서 해당 단어가 포함되고, 평가가 있는 data의 total score의 mean을 결측치의 total score로 사용하려고 했으나, 대부분 좋은 평가를 내린 항목에 대해서는 수공할 만한 수치가 도출되었으나, 중립과 나뭇잎을 표현하는 평가에 대한 수치의 평균이 높게 측정되어서 ‘그닥’이라는 단어가 포함되는 평가수치를 Raw data에서 뽑아보기로 하였다.

4.0	4.0	5.0	리뷰가 좋아 처음 주문했는데 기대했던 것보다... 불맛이 많이나지 않았고 양도 그닥...;; 리뷰음료수...
2.0	4.0	4.0	리뷰가 괜찮아서 남자친구 시켜줬는데 맛은 그닥인거 같더라고요
NaN	NaN	NaN	배송은 빨라서 좋았으나 맛은 그닥...
NaN	NaN	NaN	기대는 그닥 안했는데.. 치킨은 기대 이상으로 후라이드 일반 치킨이고 피자도 맛있네요~~ 특히도우 안고구...
2.0	2.0	3.0	여기 맛있다고 하는 애들 진짜냐? 솔직히 그닥인데 냉정하게 평가하자 시켜먹는 애들 한번 더 고민하고 시켜 먹어라 진짜
3.0	4.0	4.0	그냥저냥... 그닥 맛있지도 아주 없지도...
4.0	4.0	5.0	배달하시는 분이 그닥 친절하진 않았던 거 같습니다. 양은 많고 맛있었어요

위 그림에서 알 수 있듯, 소비자가 직접적으로 느끼는 감정은 리뷰에 작성된 개개인별로 평가하는 기준이 다르므로 상대적으로 평가하는 기준이 후한 고객도 있

는 반면, 박한 고객도 있음을 알 수 있다. 전반적으로 소비자들이 리뷰 데이터에 작성한 텍스트보다 평가를 높게 해주는 경향을 보인다. 이는 위에서 제시하였던 맛과 양에 4점, 5점을 부여한 고객이 평가한 고객 중 87%를 점유한다는 사실과 동일하게 간주할 수 있다. 따라서, 중립을 표현하는 평가의 total score에는 4.457이 아닌 3을 부여하였고, 나쁨을 표현하는 평가는 1점을 부여하였다.

.	37
?	22
o o	13
o	7
^^	7
好吃	6
1021116460	4
??	4
.	4
-	3
□ □ □	3
????	3
긱	2
□ □	2
??????????	2
제가여..여기랑성수점동시에시켰거든요?근데한쪽은양이3/4고한쪽은1	2
...	2
????????	2
????????????	2
★★★★★	2
~	2
맛	2
.....	2
긱	2
\	2
소소	2
????????	2
?好吃	2
人 人 人	2
o o o o o o o o o o	

업체명	카테고리	메뉴	맛	양	배달	리뷰	date	numbering	total	
0	전주석식 불고기-본점	한식	파절이매콤통삼겹 (2~3인) (공기밥2 + 김치찌개 + 밑반찬 + 찜) /1	5.0	5.0	5.0	자주시켜먹는단골집인데 항상 변치 않고 맛있습니 다!!!	2017년 12월 6일 수요일	0	5.0
1	전주석식 불고기-본점	한식	통삼겹살 2인 (고기 + 공기밥2 + 김치찌개 + 찜 + 밑반찬) /1(추가 선택(고기 추가))	5.0	5.0	5.0	배달시간도 오래 걸리지 않고, 양이 적을 거 같아서 고기 추가를 했는데... 안해도 됐었네요 ㅋㅋ. 맛도 이 정도...	2017년 9월 30일 토요일	1	5.0
2	전주석식 불고기-본점	한식	통삼겹살 2인 (고기 + 공기밥2 + 김치찌개 + 찜 + 밑반찬) /1	NaN	NaN	NaN	굿굿	2017년 9월 23일 토요일	2	4.5
3	전주석식 불고기-본점	한식	통삼겹 (小 / 500g) (냉얼무국수 or 냉얼무우동 or 비빔얼무국수 + 찜) /1(메뉴 선택...)	5.0	5.0	5.0	배달도 빠르고 맛나요	2018년 3월 13일 화요일	3	5.0
4	전주석식 불고기-본점	한식	통삼겹 (3~4인) (김치찌개 + 공기밥3 + 밑반찬 + 찜) /1	5.0	5.0	5.0	찌개와 통삼겹맛있는 5찬. 많이 먹는 편이라 3인분 짜리 주문했는데... 배터지는 줄 알았어요~ 단골 될 게요^^	2018년 2월 27일 화요일	4	5.0
...
28912	동강	중식	NaN	NaN	NaN	NaN	좋아요	2013년 12월 31일 화요일	28912	4.5
28913	동강	중식	NaN	NaN	NaN	NaN	맛있ㅇ ㅜㅜ	2013년 12월 29일 일요일	28913	4.5
28914	동강	중식	NaN	NaN	NaN	NaN	맛있게 다	2013년 12월 25일 수요일	28914	4.5
28915	동강	중식	NaN	NaN	NaN	NaN	여기 맛있어요 이 근처 탕수육은 냄새? ㅜㅜ 돼지 냄새? 나는 곳이너무 많은데 여긴 안그래요 짜장도 맛있구 찜통...	2013년 12월 10일 화요일	28915	4.5
28916	동강	중식	NaN	NaN	NaN	NaN	맛있어요 또 시켜먹어야지 추천할게요	2013년 11월 13일 수요일	28916	4.5

맛, 양, 배달 항목의 결측치가 존재해도 total score가 전부 잘 산정됨을 확인할 수 있었으며, 정제 후 data의 shape은 (27858, 10)이다. 해당 데이터의 리뷰 항목에서 띄어쓰기를 전부 제거하였으므로, 기존의 Raw data와 numbering을 기준으로 total column과 numbering column을 inner join을 하여 리뷰 항목의 텍스트를 복구하였다.


```
df_fin = pd.merge(df_origin_ver, df_numbering_total, left_on = 'numbering', right_on = 'numbering', how = 'inner')
df_fin
```

	업체명	카테고리	메뉴	맛	양	배달	리뷰	date	numbering	total
0	전주석식물고기-본점	한식	파절이매콤통삼겹 (2~3인) (공기밥2 + 김치찌개 + 밀반찬 + 찜) /1	5.0	5.0	5.0	자주시켜먹는 단골집인데 항상변치않고 맛있습니다!!	2017년 12월 6일 수요일	0	5.0
1	전주석식물고기-본점	한식	통삼겹살 2인 (고기 + 공기밥2 + 김치찌개 + 찜 + 밀반찬) /1(추가 선택(고기 추가))	5.0	5.0	5.0	배달 시간도 오래걸리지 않고, 양이 적을거 같아서 고기 추가를 했는데..안해도 됐었...	2017년 9월 30일 토요일	1	5.0
2	전주석식물고기-본점	한식	통삼겹살 2인 (고기 + 공기밥2 + 김치찌개 + 찜 + 밀반찬) /1	NaN	NaN	NaN	굿굿	2017년 9월 23일 토요일	2	4.5
3	전주석식물고기-본점	한식	통삼겹 (小 / 500g) (냉얼무국수 or 냉얼무우동 or 비빔얼무국수 + 찜) /1(메뉴 선택...)	5.0	5.0	5.0	배달도빠르고맛나요	2018년 3월 13일 화요일	3	5.0
4	전주석식물고기-본점	한식	통삼겹 (3~4인) (김치찌개 + 공기밥3 + 밀반찬 + 찜) /1	5.0	5.0	5.0	찌개와 통삼겹 맛있는 5찬. 많이 먹는편이라 3인분짜리 주문했는데... 배터지는줄 ...	2018년 2월 27일 화요일	4	5.0
...
27853	동강	중식		NaN	NaN	NaN	좋아요	2013년 12월 31일 화요일	28912	4.5
27854	동강	중식		NaN	NaN	NaN	맛있ㅇㅏ웃	2013년 12월 29일 일요일	28913	4.5
27855	동강	중식		NaN	NaN	NaN	맛이게다	2013년 12월 25일 수요일	28914	4.5
27856	동강	중식		NaN	NaN	NaN	여기 맛있어요 이근쳐 탕수육은 냄새?ㅈㅈ돼지냄새?나는 곳이 너무 많은데 여긴안그래요...	2013년 12월 10일 화요일	28915	4.5
27857	동강	중식		NaN	NaN	NaN	맛있어요 또시켜먹어야지 추천할게요	2013년 11월 13일 수요일	28916	4.5

‘리뷰’ column에 띄어쓰기가 복구됨과 동시에 total score가 정상적으로 생긴 데이터프레임을 확인할 수 있다. 우리는 감성분석이 주 목적이므로, 위 데이터프레임에서 리뷰 column과 total column을 빼고, total column의 이름을 rating으로 rename한다.

	리뷰	rating
0	자주시켜먹는 단골집인데 항상변치않고 맛있습니다!!	5.0
1	배달 시간도 오래걸리지 않고, 양이 적을거 같아서 고기 추가를 했는데..안해도 됐었...	5.0
2	굿굿	4.5
3	배달도빠르고맛나요	5.0
4	찌개와 통삼겹 맛있는 5찬. 많이 먹는편이라 3인분짜리 주문했는데... 배터지는줄 ...	5.0
...
27853	좋아요	4.5
27854	맛있ㅇㅏ웃	4.5
27855	맛이게다	4.5
27856	여기 맛있어요 이근쳐 탕수육은 냄새?ㅈㅈ돼지냄새?나는 곳이 너무 많은데 여긴안그래요...	4.5
27857	맛있어요 또시켜먹어야지 추천할게요	4.5

27858 rows x 2 columns

위 Dataset의 shape은 (27858, 2)이며, null값은 없음을 확인할 수 있다.

```
#information
df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 27858 entries, 0 to 27857
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  -
0   리뷰    27858 non-null    object
1   rating  27858 non-null    float64
dtypes: float64(1), object(1)
memory usage: 652.9+ KB
```

```
# 결측치
df.isnull().sum()

리뷰      0
rating    0
dtype: int64
```

1차적으로 정제된 Dataset에서도 추가적으로 정제해야 할 사항을 발견하였다. 리뷰 data의 value로 특수문자, 모음이 존재하는 경우를 확인할 수 있었다. 이는 감성분석 및 Text mining을 적용할 의미가 없으므로, 정규표현식을 이용하여 제거 하도록 한다.

	리뷰	rating			
	존맛★	3.666667	ㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇ	4.500000	
	!는데 굽★볼케이노보다 더매...	4.500000	ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ	4.500000	
	은 강민★입니다. 브랜드로 손...	4.500000	ㅎㅎ 닭도 부드럽고 닭...	5.000000	
	주 먹을 것 같아요 ★ 맛있네요!	4.500000	ㅇㅇㅇㅇㅇㅇ 트ㅏㅏㅏㅇ	3.000000	
	죽은 ★★★★★ 친절 ★★★★★	4.500000	매우면서 달달해요 맛...	4.500000	
	ㅠㅠ 군만두먹고싶었는데.제....	4.500000	맛있ㄴㄱ다 스테이크ㅇㅇ	5.000000	
	달원분이 오분후 도착이라고 ...	4.000000	맛먹기좋아요 ㅇ지척ㅇ	4.500000	
	게 마늘소스 부어서 먹으면 짭...	5.000000	도 너무너ㅏㅇ 좋은곳!!!!	5.000000	
	하고 너무맛있었어요!! 강추★★	4.500000	먹다보니살짝매콤함이^^	4.000000	
	귀찮아도 매번 리뷰남겨요ㅎ ...	4.500000	. 맛있ㅇㅇㅇ배달도 굳굳	5.000000	
	도 강추해서 주문했는데 진짜 ...	3.666667	ㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇ	5.000000	
	★★★★★?	3.333333	ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ	5.000000	
			ㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇ	5.000000	
			ㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇ	5.000000	
			ㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇ	5.000000	
			이 많이 불어터짐ㅇㅇ...	4.500000	

특수문자와 모음이 리뷰 data에 고스란히 남아있음을 확인할 수 있다.

3-3) Korean Text data Preprocessing

기계가 텍스트 형식으로 되어있는 리뷰 데이터를 이해하기 위해서, 텍스트 데이터를 단어 단위로 분리하는 전처리 과정이 필요하다. 여기서 분리된 단어들은 Bag of Words count 기반으로 표현할 수 있고, TF-IDF를 사용해 수치로 나타낼 수도 있다. 리뷰의 내용을 단어화하여 형태소를 추출하고 Bag of Words를 생성하여 TF-IDF 변환을 진행한다.

```
import re

def apply_regular_expression(text):
    hangul = re.compile('[^ㄱ-ㅣ 가-힣]') # 한글 추출 규칙: 띄어 쓰기(1 개)를 포함한 한글
    result = hangul.sub('', text) # 위에 설정한 "hangul" 규칙을 "text"에 적용(.sub)시킴
    return result
```

```
df['리뷰'][20869]
```

```
'우오...ㅇㅇㅁㅁ워요ㅋㅋ 원래 매운거 조아하는데 .. 먹다가 땀났슈ㅎㅎ 닭도 부드럽고 닭발도 좋겟? 근디 먹다가 두개가 빼 덜발라졌슈ㅎㅎ'
```

```
apply_regular_expression(df['리뷰'][20869])
```

```
'우오ㅁㅁㅁ워요ㅋㅋ 원래 매운거 조아하는데 먹다가 땀났슈ㅎㅎ 닭도 부드럽고 닭발도 좋겟 근디 먹다가 두개가 빼 덜발라졌슈ㅎㅎ'
```

정규표현식을 적용한 후 특수문자가 잘 제거됨을 확인할 수 있다.

```
[ ] !pip install konlpy
    from konlpy.tag import Okt
    from collections import Counter
    !pip install --upgrade pip
    !pip install konlpy
```

```
[ ] okt = Okt() # 명사 형태소 추출 함수
    nouns = okt.nouns(apply_regular_expression(df['리뷰'][20869]))
    nouns
```

['오', '워', '원래', '땀났슈', '닭', '닭발', '꿀깃', '디', '개', '뼈', '덜', '슈']

명사 형태소 추출 함수를 사용하여 정규표현식을 적용한 위 리뷰 내용의 형태소를 추출하였다. 이를 전체 말뭉치(corpus)에 적용하여, 명사 형태소를 추출한다.

```
# 말뭉치 생성
corpus = "".join(df['리뷰'].tolist())
corpus
```

☞ '자주시켜먹는 단골집인데 항상변치않고 맛있습니다 !!배달 시간도 오래걸리지 않고, 양이 적을거 같아서 많이 먹는편이라 3인분짜리 주문했는데... 배터지는줄 알았어요~ 단골될께요^^일인분같지않은양.. 냉면안시 겹살이 조금 느끼한거같긴한데 전체적으로 맛있어요?맛나용 고기가 두꺼워서 좋아요 ㅎㅎ 양이 조금 더 있었 고기양은많아요 맛나용괜찬네요.양도 많고 맛있었어요 ! 고기 양념도 달달해서 맛있고 갈비만두가 속이 꽉차 빔국수로 가져왔네요와우... 간만에 시켰는데 정말맛있네요! 시원한냉면. 푸짐하고 맛있는고기! 술안주와 밥 면육수만ㅎ 어쨌든 저희는 잘먹었습니다. 번창하세요!!!잘먹었습니다 물냉도 얼음이랑 육수 많이들오기 다음에도 많이주세요맛도 좋고 배달도 빨라요비냉 맛남니다 잘먹었어요맛났었어요 ㅎㅎㅎㅎㅎ마싯섯서영 ㅎㅎ 밥이랑 찌개시켜서 먹음 편창을듯도 해요. 음식은 간만 빼면 깔끔했습니다그냥 그래욤....고기는 맛있어요

```
[ ] apply_regular_expression(corpus)
```

'자주시켜먹는 단골집인데 항상변치않고 맛있습니다 배달 시간도 오래걸리지 않고 양이 적을거 같아서 고기 편이라 인분짜리 주문했는데 배터지는줄 알았어요 단골될께요일인분같지않은양 냉면안시켰음배고팠을뻔 편다 같긴한데 전체적으로 맛있어요맛나용 고기가 두꺼워서 좋아요 ㅎㅎ 양이 조금 더 있었으면 하는 아쉬움이 있 찬네요양도 많고 맛있었어요 고기 양념도 달달해서 맛있고 갈비만두가 속이 꽉차있는데 너무 맛있어요 ㅎㅎ 시켰는데 정말맛있네요 시원한냉면 푸짐하고 맛있는고기 술안주와 밥반찬을 동시에해결ㅋ 자알먹었습니다 디 니다 물냉도 얼음이랑 육수 많이들오가있어서 시원하게 먹고 고기도맛있네요가끔시켜먹는데 정말 먹음만하 어용 ㅎㅎㅎㅎㅎ마싯섯서영 ㅎㅎㅎㅎㅎ양은 무척 많네요 근데 너무 짜요 냉면 육수도 짜고 고기도 짜서 같이 고기는 맛있어요 비냉시켰는데 물냉와서 다시기다리느라 오래기달렸네요양념고기가참맛나네요 물냉은 상상하

이후, 전체 말뭉치에서 명사 형태소를 추출한다.

```
[ ] # 전체 말뭉치(corpus)에서 명사 형태소 추출
nouns = okt.nouns(apply_regular_expression(corpus))
print(nouns)
```

['자주', '단골', '집', '항상', '배달', '시간', '양', '고기', '추가', '안해', '맛', '정도', '생각', '습니', '곳곳',

명사 형태소 추출 전 문장

자주시켜먹는 단골집인데 항상변치않고
맛있습니다 배달 시간도 오래걸리지 않
고 양이 적을거 같아서 고기 ...

추출된 명사 형태소

‘자주’, ‘단골’, ‘집’, ‘항상’, ‘배달’, ‘시
간’, ‘양’, ‘고기’ ...

```
[ ] counter = Counter(nouns)
```

```
[ ] counter.most_common(10)
```

```
[('맛', 8939),
 ('배달', 6132),
 ('족발', 2193),
 ('진짜', 2004),
 ('주문', 1937),
 ('좀', 1798),
 ('양도', 1716),
 ('치킨', 1698),
 ('정말', 1631),
 ('또', 1425)]
```

왼쪽의 결과에서 볼 수 있듯, 두 글자
키워드가 대부분이 유의미한 단어이지
만, ‘맛’, ‘좀’, ‘또’와 같은 한 글자 키워
드는 분석에 딱히 좋은 영향을 미치지
못할 것으로 판단된다. ‘맛’도 ‘맛있다’
와 ‘맛없다’의 이중적인 해석이 가능하
다. 따라서 한 글자 명사는 분석에서
배제하도록 한다.

```
available_counter = Counter({x: counter[x] for x in counter if len(x) > 1})
available_counter.most_common(50)
```

```
[('배달', 6132),
 ('족발', 2193),
 ('진짜', 2004),
 ('주문', 1937),
 ('양도', 1716),
 ('치킨', 1698),
 ('정말', 1631),
 ('항상', 1397),
 ('시간', 1296),
 ('여기', 1266),
 ('피자', 1228),
 ('리뷰', 991),
 ('최고', 963),
 ('소스', 946),
 ('튀김', 933),
 ('처음', 929),
 ('달걀', 922),
 ('다음', 919),
 ('자주', 884),
 ('서비스', 854),
 ('역시', 850),
 ('막국수', 833),
 ('조금', 821),
 ('가격', 819),
 ('완전', 798),
 ('생각', 786),
 ('양념', 732),
 ('오늘', 730),
```

한 글자 키워드를 전부 제거했지만, ‘진짜’, ‘항상’과 같은 실질적인 의미가 없는 꾸밈의 역할을 하는 불용어들이 아직 존재한다. 한국어 불용어 사전을 정의하여 불용어또한 제거하도록 한다.

```
stopwords = pd.read_csv("https://raw.githubusercontent.com/yonkt200/FastCampusDataset/master/korean_stopwords.txt").values.tolist()
stopwords[:20]
```

```
[['휴'],
 ['아이구'],
 ['아이쿠'],
 ['아이고'],
 ['어'],
 ['나'],
 ['우리'],
 ['저희'],
 ['따라'],
 ['의해'],
 ['을'],
 ['를'],
 ['에'],
 ['의'],
 ['가']]
```

이외에도, 우리가 분석하고자 하는 리뷰 데이터셋에 특화된 불용어들이 존재한다. 위에서 볼 수 있듯 ‘족발’, ‘치킨’과 같은 음식 이름은 긍정과 부정을 판단하기에 어려움이 있는 단어이다. 따라서 count가 높게 잡힌 단어들을 불용어 사전에 추가한다.

```
[ ] yogiyo_stopwords = ['족발', '치킨', '피자', '닭발', '막국수', '떡볶이', '탕수육', '주먹밥', '고기', '짬뽕', '보쌈']
for word in yogiyo_stopwords:
    stopwords.append(word)
```

3-4) Word Count

Bow 벡터를 생성하는 과정은 다음과 같다.

```
[ ] from sklearn.feature_extraction.text import CountVectorizer

def text_cleaning(text):
    hangul = re.compile('[^ㄱ-ㅣ가-힣]') # 정규 표현식 처리
    result = hangul.sub('', text)
    okt = Okt() # 형태소 추출
    nouns = okt.nouns(result)
    nouns = [x for x in nouns if len(x) > 1] # 한글자 키워드 제거
    nouns = [x for x in nouns if x not in stopwords] # 불용어 제거
    return nouns

vect = CountVectorizer(tokenizer = lambda x: text_cleaning(x))
bow_vect = vect.fit_transform(df['리뷰']).tolist()
word_list = vect.get_feature_names()
count_list = bow_vect.toarray().sum(axis=0)
```

```
[ ] count_list #각 단어가 전체 리뷰중에 등장한 총 횟수

array([ 1,  1, 63, ...,  1,  1,  1])
```

```
[ ] # 각 단어의 리뷰별 등장 횟수
bow_vect.toarray()

array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]])
```

```
[ ] bow_vect.shape

(27858, 7898)
```

3-5) TF-IDF 적용

생성된 Bag of Words 벡터에 대해 TF-IDF 변환을 진행한다.

```
[ ] from sklearn.feature_extraction.text import TfidfTransformer

tfidf_vectorizer = TfidfTransformer()
tf_idf_vect = tfidf_vectorizer.fit_transform(bow_vect)
```

```
[ ] print(tf_idf_vect.shape)

(27858, 7898)
```

```
[ ] # 첫 번째 리뷰에서의 단어 중요도(TF-IDF 값) -- 0이 아닌 것만 출력
print(tf_idf_vect[0])
```


```
(0, 7570)    0.4870508989008128
(0, 5760)    0.5375292554312999
(0, 1396)    0.6883630738464859
```

변환 후, 27858 x 7898 행렬이 출력된다. 여기서 한 행(row; 27,858)은 리뷰의 개수를 의미하고, 한 열(column; 7898)은 단어의 개수를 의미한다.

```
[ ] # 첫 번째 리뷰에서 모든 단어의 중요도 -- 0인 값까지 포함
print(tf_idf_vect[0].toarray().shape)
print(tf_idf_vect[0].toarray())

(1, 7898)
[[0. 0. 0. ... 0. 0. 0.]]
```

벡터와 단어의 mapping 작업을 실행해준다.

 vect.vocabulary_

```
{'자주': 5760,
 '단골': 1396,
 '항상': 7570,
 '배달': 3041,
 '시간': 4149,
 '추가': 6824,
 '안해': 4527,
 '정도': 5995,
 '생각': 3780,
 '습니': 1831,
 '굿굿': 701,
 '찌개': 6631,
 '삼겹': 3707,
 '편이': 7301,
 '인분': 5597,
 '주문': 6234,
 '냉면': 1143,
 '안시': 4494,
 '마늘': 2126,
```

```
[ ] invert_index_vectorizer = {v: k for k, v in vect.vocabulary_.items()}
print(str(invert_index_vectorizer)[:100]+'...')
```

```
{5760: '자주', 1396: '단골', 7570: '항상', 3041: '배달', 4149: '시간', 6824: '추가
```

벡터와 단어의 mapping 작업이 정상적으로 되었음을 확인할 수 있다.

3-5) Logistic Regression을 이용한 감성 분류 작업

전처리된 리뷰 데이터를 활용하여 감성 분류 예측 모델을 생성한다. 감성 분류 예측 모델이란, 소비자의 리뷰 평가 내용을 통해서 해당 리뷰가 긍정적인지, 혹은 부정적인지 예측하여 이용자의 감성을 파악하는 것이다. 따라서, 모델의 X값(feature의 값)은 소비자 리뷰의 평가 내용, 즉 리뷰 text data가 되는 것이고, 모델의 Y값(label 값)은 이용자의 긍정 혹은 부정적인 감성을 의미한다.

이용자의 리뷰를 ‘긍정’과 ‘부정’ 두 가지로 이진분류를 하고자 한다. 하지만 이용

자의 감성을 대표할 수 있는 평가점수, 즉 rating 변수는 1부터 5의 수치를 가지고 있다. 따라서 평가 점수 변수를 이진변수로 변환해야한다.

df.sample(10)

	리뷰	rating
29	맛있고 배달빠르고 전화응대도 친절하네요 고기 매운맛보다 석쇠를 추천합니다	5.000000
4966	너무너무 맛나요^__^	4.500000
18586	배달이 엄청 빨라요 양도많고!	4.500000
14482	뒷발이 정말 존독존독하고 살도많고 앞으론도야족발에서만시켜먹어야겠네요.. 잘먹었습니다!!	4.500000
14773	맛있게 잘 먹었습니다.	4.666667
1364	개꿀맛 피자 꼭 먹으셈	4.500000
27440	량이 많아서 좋아요	4.500000
5025	맛도좋구양도많구 잘먹었네요	4.500000
11545	90분을 기다려서 먹은건 거의 다식어서 온 치킨이었네요. 만든지 시간이 지나서 습기랑...	4.500000
10768	맛있어요~배달도 늦지않았어요	4.333333

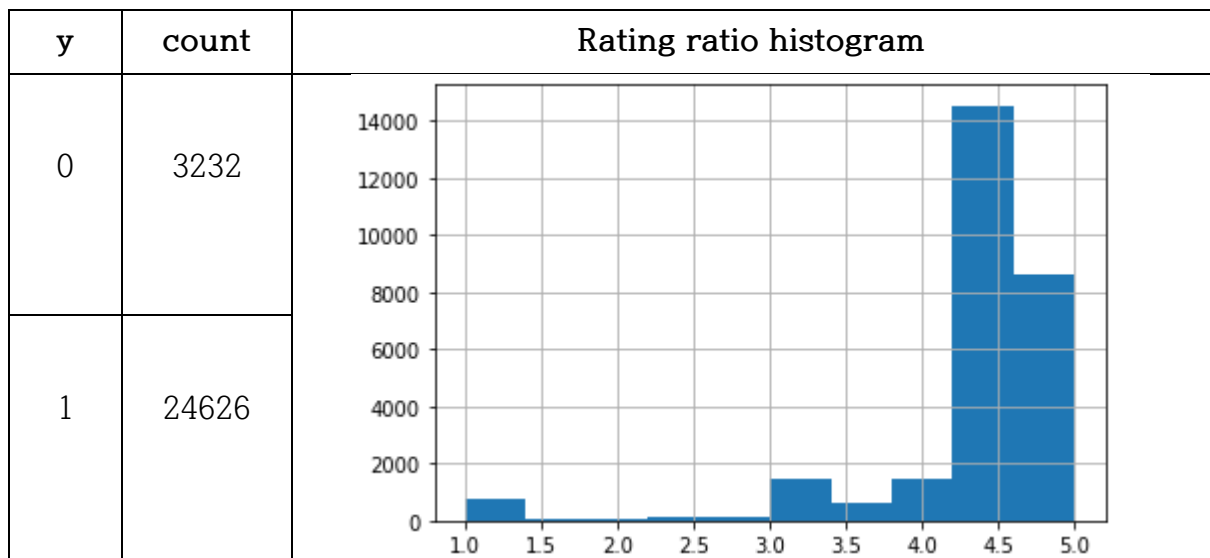
리뷰의 전반적인 내용과 평점을 살펴보면, 4~5점의 rating을 가진 리뷰는 대부분 긍정적이었지만, 1~3점의 리뷰에서는 부정적인 평가와 중립적인 평가를 보였다. 중립적인 평가에 대하여 어떻게 분류를 해야 할지에 대하여 고민하였는데, 크게 ‘맛있다’와 ‘아니다’로 분류해야 한다고 생각했기에 긍정적인 리뷰는 4점 이상의 rating을 가진 데이터로 분류하여 label값에 1을 부여하고, 4점 미만의 리뷰는 부정적인 리뷰로 분류하여 0을 부여하였다.

```
[ ] def rating_to_label(rating):
    if rating >= 4:
        return 1
    else:
        return 0

df['y'] = df['rating'].apply(lambda x: rating_to_label(x))
```

df.sample(10)

	리뷰	rating	y
2854	배달빠르고 맛도 좋았어요^^	4.5	1
25029	배달아저씨친절 요리는평균이상 배달음식쓰레기많은데 여기는괜찮았음	4.5	1
3496	잘먹었습니다 덕분에 두끼 해결했네요	5.0	1
5935	엄청 느끼해요 배달도 너무 느리고ㅠㅠ 식히고 먹으니깐 그래도 그나마 먹을만했어요	5.0	1
7093	만족만족 맛있습니다	4.5	1
10624	새콤,매콤,달콤 너무 맛있게 잘먹었어요. 사업번창하세요	4.5	1
14893	배달이 빨리되서 오래 기다리지 않고 맛있게 잘먹었어요~ 주먹밥도 오고 매운족발도 적...	4.5	1
22110	맛있고요 배달빨랐어요 늦을줄알고 나갔다가 급전화드려서 집에 있는 사람한테 결제요청했...	4.5	1
13389	짜장면은괜찮구요 짬뽕은싱겁네요!	3.0	0
16510	배달이참빠르고 맛있게먹는중입니당	4.5	1



이후, 모델 학습을 위해서 training set과 test set을 분류하였다.

```
[ ] from sklearn.model_selection import train_test_split

x = tf_idf_vect
y = df['y']
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state=1)
```

```
[ ] x_train.shape, y_train.shape

((19500, 7898), (19500,))
```

```
[ ] x_test.shape, y_test.shape

((8358, 7898), (8358,))
```

```
▶ from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# fit in training set
lr = LogisticRegression(random_state = 0)
lr.fit(x_train, y_train)

# predict in test set
y_pred = lr.predict(x_test)
```

```
[ ] # classification result for test set

print('accuracy: %.2f' % accuracy_score(y_test, y_pred))
print('precision: %.2f' % precision_score(y_test, y_pred))
print('recall: %.2f' % recall_score(y_test, y_pred))
print('F1: %.2f' % f1_score(y_test, y_pred))

accuracy: 0.89
precision: 0.89
recall: 1.00
F1: 0.94
```

Logistic Regression 모델의 학습 결과

Accuracy, precision, recall, f1 score 모두 높은 수치를 보였다. Confusion matrix 를 통해서 예측이 올바르게 되었는지 분류 모델을 확인한다.

일반적인 Confusion Matrix 의 형태는 다음과 같다.

		Predicted class	
		<i>P</i>	<i>N</i>
Actual Class	<i>P</i>	True Positives (TP)	False Negatives (FN)
	<i>N</i>	False Positives (FP)	True Negatives (TN)

TP: 관심 범주를 정확하게 분류한 값

FN: 관심 범주를 관심 범주가 아닌

것으로 잘못 분류한 값

FP: 관심 범주가 아닌 값을 관심

범주로 잘못 분류한 값

TN: 관심 범주가 아닌 것을 정확하게

분류한 값

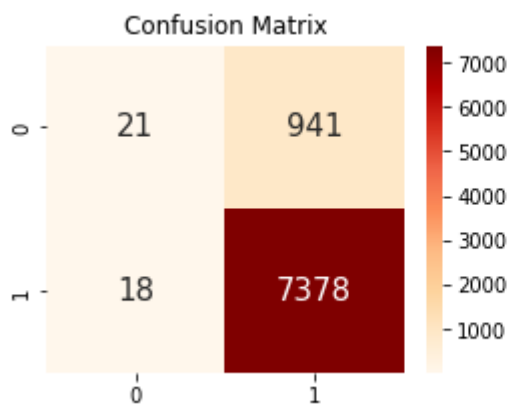
평가 척도	Accuracy (정확도)	Precision (정밀도)	Recall (재현도)	F1 - score
	$\frac{TP + TN}{TP + TN + FP + FN}$	$\frac{TP}{TP + FP}$	$\frac{TP}{TP + FN}$	$\frac{Precision * Recall}{Precision + Recall}$

```
[ ] # confusion matrix

from sklearn.metrics import confusion_matrix

confu = confusion_matrix(y_true = y_test, y_pred = y_pred)

plt.figure(figsize=(4, 3))
sns.heatmap(confu, annot=True, annot_kws={'size':15}, cmap='OrRd', fmt='.10g')
plt.title('Confusion Matrix')
plt.show()
```



모델의 평가 결과를 살펴보면, 모델이 지나치게 긍정적으로만 ($y = '1'$) 예측하려는 경향이 있음을 확인할 수 있었다. 이는 26p 에 제시한 히스토그램에서 문제를 확인할 수 있었다. 애초에 샘플 데이터의 클래스가 1 인 경우가 0 인 경우보다 매우 많기에, 지나치게 1 로 예측하려는 경향을 보였다. 즉, 샘플 데이터의 클래스 불균형으로 인한 문제로 해석할 수 있다. 클래스 불균형을 1:1 샘플링을 통해서 재조정하고자 한다.

```
[ ] df['y'].value_counts()
```

```
1    24626
0     3232
Name: y, dtype: int64
```

```
[ ] positive_random_idx = df[df['y']==1].sample(3232, random_state=12).index.tolist()
negative_random_idx = df[df['y']==0].sample(3232, random_state=12).index.tolist()
```

```
[ ] random_idx = positive_random_idx + negative_random_idx
x = tf_idf_vect[random_idx]
y = df['y'][random_idx]
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state=1)
```

```
[ ] x_train.shape, y_train.shape
```

```
((4848, 7898), (4848,))
```

```
[ ] x_test.shape, y_test.shape
```

```
((1616, 7898), (1616,))
```

클래스를 1:1 샘플링을 통하여 재조정하였고, 모델을 재학습한다.

```
[ ] lr2 = LogisticRegression(random_state = 0)
lr2.fit(x_train, y_train)
y_pred = lr2.predict(x_test)

[ ] # classification result for test set

print('accuracy: %.2f' % accuracy_score(y_test, y_pred))
print('precision: %.2f' % precision_score(y_test, y_pred))
print('recall: %.2f' % recall_score(y_test, y_pred))
print('F1: %.2f' % f1_score(y_test, y_pred))
```

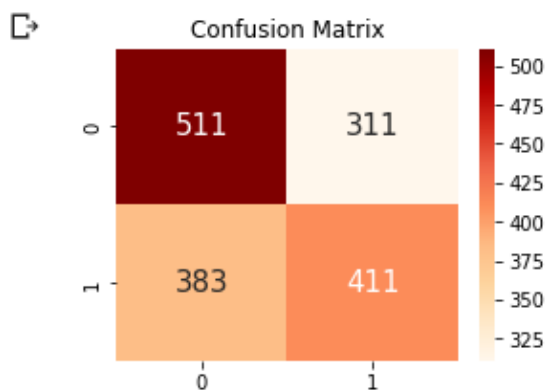
```
accuracy: 0.57
precision: 0.57
recall: 0.52
F1: 0.54
```

```
# confusion matrix

from sklearn.metrics import confusion_matrix

confu = confusion_matrix(y_true = y_test, y_pred = y_pred)

plt.figure(figsize=(4, 3))
sns.heatmap(confu, annot=True, annot_kws={'size':15}, cmap='OrRd', fmt='.10g')
plt.title('Confusion Matrix')
plt.show()
```



Label 이 0 인 샘플과 1 인 샘플의 개수를 동등하게 하는 1:1 sampling 을 진행하였으나, 평가 척도(Accuracy, Precision, Recall, F1 Score)는 더 낮아졌다. 그러나 이전과 같이 지나치게 1 로 예측하려는 경향은 피할 수 있었다. 앞에서 언급했듯, 기본적으로 고객이 서비스에 대하여 불만족감을 느꼈어도 평가는 본인이

느낀 감정에 비해 후하게 주는 경우가 빈번했다. 그로 인하여 Raw data 의 ‘맛’, ‘양’, ‘배달’의 평가 척도 또한 대부분 4 점과 5 점으로 평가되었기에 해당 오류가 발생하였다고 보여진다.

3-6) 긍정 및 부정 키워드 분석

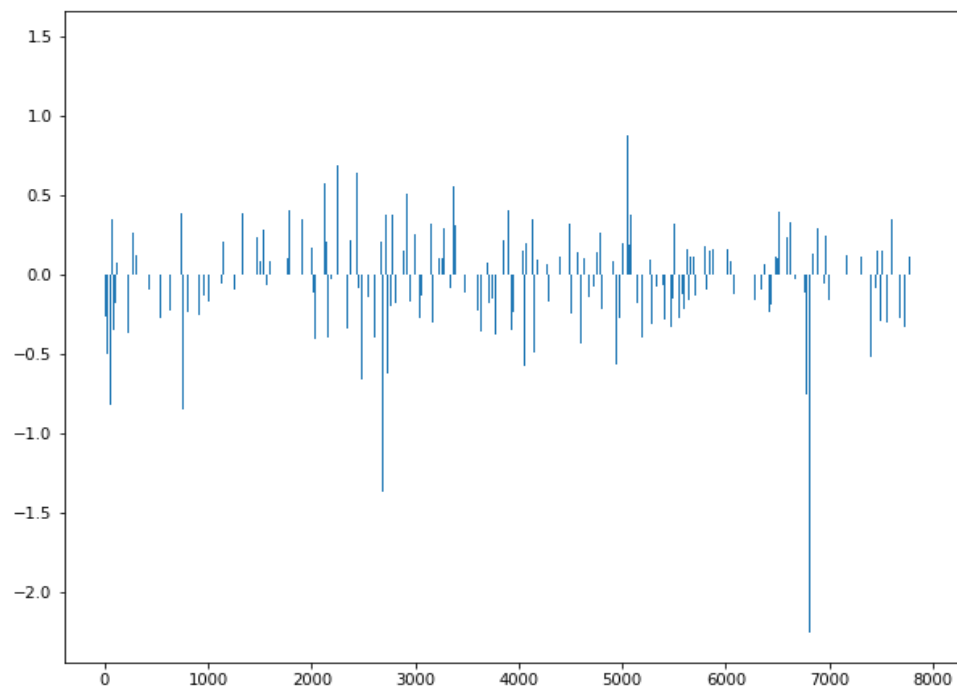
학습된 Logistic Regression 모델을 이용해, 긍정 및 부정 키워드를 추출해낼 수 있다. 추출된 키워드를 통하여 이용자가 느끼는 요식업체, 음식 및 서비스의 장, 단점을 파악할 수 있고 이를 기반으로 앞으로 유지해야 할 좋은 서비스와, 개선이 필요한 아쉬운 서비스에 대해서 어느정도 판단할 수 있는 근거를 마련할 수 있다. 키워드를 추출하기 위해, Logistic Regression 모델의 각 단어의 coefficient 를 시각화한다.

```
[ ] lr2.coef_
```

```
array([[0.20613607, 0.          , 0.96304168, ..., 0.          , 0.          ,  
       0.          ]])
```

```
# print logistic regression's coef  
plt.figure(figsize=(10, 8))  
plt.bar(range(len(lr2.coef_[0])), lr2.coef_[0])
```

<BarContainer object of 7898 artists>



위 그림에서 계수, 즉 Coefficient 가 양인 경우는 단어가 긍정적인 영향을 미쳤다고 판단할 수 있다. 반면, 계수가 음인 경우는 자연스럽게 부정적인 영향을 미쳤다고 판단할 수 있다. 해당 계수들을 크기순으로 정렬하여, 긍정 및 부정 키워드를 출력한다.

```
[ ] print(sorted(((value, index) for index, value in enumerate(lr2.coef_[0])), reverse = True)[:5])
print(sorted(((value, index) for index, value in enumerate(lr2.coef_[0])), reverse = True)[-5:])
# enumerate: 인덱스 번호와 컬렉션의 원소를 tuple형태로 반환함

[(1.4747933415266463, 1526), (1.4702771444651026, 466), (1.3564783360458843, 4898), (1.3451456859434676, 2420), (1.2736136633330337, 3131)]
[(-1.610608473974656, 3239), (-1.6823145121162713, 4296), (-1.8751004612758821, 2261), (-1.932627581813851, 749), (-2.254240350886336, 6812)]
```

긍정 키워드와 부정 키워드의 상위 5 개를 출력하였다.

```
coef_pos_index = sorted(((value, index) for index, value in enumerate(lr2.coef_[0])), reverse = True)
coef_neg_index = sorted(((value, index) for index, value in enumerate(lr2.coef_[0])), reverse = False)
coef_pos_index

[(1.4747933415266463, 1526),
 (1.4702771444651026, 466),
 (1.3564783360458843, 4898),
 (1.3451456859434676, 2420),
 (1.2736136633330337, 3131),
 (1.1993258838903593, 5883),
 (1.195040346588146, 4113),
 (1.1824299466902637, 4988),
```

전체 단어가 포함된 긍정 키워드 리스트와 부정 키워드 리스트를 정의하고 출력한다.

```
invert_index_vectorizer = {v: k for k, v in vect.vocabulary_.items()}
invert_index_vectorizer

{5760: '자주',
 1396: '단골',
 7570: '항상',
 3041: '배달',
 4149: '시간',
 6824: '추가',
 4527: '안해',
 5995: '정도',
```


마지막으로, index 를 단어로 변환하여 긍정 키워드 리스트와 부정 키워드 리스트의 top 20 를 출력한다.

```
for coef in coef_pos_index[:20]:
    print(invert_index_vectorizer[coef[1]], coef[0])
```

```
대박 1.4747933415266463
고추 1.4702771444651026
엽떡 1.3564783360458843
맛집 1.3451456859434676
번창 1.2736136633330337
저녁 1.1993258838903593
스트레스 1.195040346588146
오시 1.1824299466902637
단골 1.1503889125535784
이틀 1.1445034578918194
베리 1.1143205348435836
치즈 1.1115677241805844
날치 1.1009570774827575
홍닭 1.0696935878093052
아저씨 1.020595416960358
짱짱 1.0202118184948883
이용 1.0143522993193934
불닭 1.0042322504587993
야들야들 0.9805514340469808
걱정 0.9800222241062875
```

```
for coef in coef_neg_index[:20]:
    print(invert_index_vectorizer[coef[1]], coef[0])
```

```
최악 -2.254240350886336
그냥 -1.932627581813851
만해 -1.8751004612758821
실망 -1.6823145121162713
보통 -1.610608473974656
한지 -1.3686869214968318
무난 -1.3632621248623875
예정 -1.3579753305404492
볶음밥 -1.3558836089975894
별로 -1.3387918097579923
동강 -1.3304171498171553
시간 -1.2366516688019782
그거 -1.177183326639892
김치찌개 -1.1388382966903419
비빔밥 -1.0801158436554215
괜춘 -1.0667141238598972
불고기 -1.0634387580414435
취소 -1.0511937413302357
대비 -1.021861413207081
단무지 -0.993087849603095
```

긍정 키워드와 부정 키워드를 정리해보면 다음과 같다.

긍정 키워드	부정 키워드
대박, 고추, 엽떡, 맛집, 번창, 저녁, 스트레스, 오시, 단골, 이틀, 베리, 치즈, 날치, 홍닭, 아저씨, 짱짱, 이용, 불닭, 야들야들, 걱정	최악, 그냥, 만해, 실망, 보통, 한지, 무난, 예정, 볶음밥, 별로, 시간, 그거, 김치찌개, 비빔밥, 괜춘, 불고기, 취소, 대비, 단무지

IV. 분석 결과 해석

여러 기법들을 사용해서, 제시된 데이터의 긍정 키워드와 부정 키워드는 다음과 같이 정리할 수 있다.

긍정 키워드	부정 키워드
대박, 고추, 엽떡, 맛집, 번창, 저녁, 스트레스, 오시, 단골, 이틀, 베리, 치즈, 날치, 홍담, 아저씨, 짱짱, 이용, 불담, 야들야들, 걱정	최악, 그냥, 만해, 실망, 보통, 한지, 무난, 예정, 볶음밥, 별로, 시간, 그거, 김치찌개, 비빔밥, 팬션, 불고기, 취소, 대비, 단무지

여기서, 처음 쓰여진 단어들이 가장 긍정적인 혹은 부정적인 단어를 나타내며, 나중에 쓰여 질수록 앞서 쓰여진 단어들에 비해 상대적으로 덜 긍정적인 혹은 덜 부정적인 단어임을 의미한다.

4-1) 긍정 키워드 분석

일반적으로 ‘대박’, ‘맛집’, ‘번창’, ‘단골’, ‘짱짱’, ‘야들야들’ 이라는 단어는 긍정적인 의미로 받아들일 수 있다. 그러나 ‘고추’, ‘엽떡’, ‘저녁’, ‘스트레스’, ‘오시’, ‘이틀’, ‘베리’, ‘아저씨’, ‘이용’, ‘걱정’이라는 단어를 보면 바로 긍정적인 키워드라고 유추할 수 없다. 해당 단어가 포함되어있는 리뷰 data를 개별적으로 관찰하여 해석하고자 한다.

단어	주 점유 업체	대표 문장	해석
고추	굽네치킨-화양점 전주석쇠불고기-본점 직화신불담	고추장으로 먹었는데 맵지만 맛있어요! 잘먹었습니다~, 고추장도 맛나지만 매운거 못드시는분들은 간장 강추^^♡♡, 배달도 빠르고 치킨도 역시나 ♡? 고추바사삭 사랑해요 굽네	굽네치킨의 ‘고추바사삭’ 메뉴, 전주석쇠불고기의 ‘고추장’이 들어간 메뉴가 소비자들에게 높은 만족을 보이고 있다. 고추라는 단어는 주로 매운

			메뉴의 리뷰에 작성되었는데, '맛있게 맵다', '스트레스가 풀린다'라는 반응이 많았다.
엽떡	불난떡볶이-광진구점 짬뽕-통오징어떡볶이	엽떡보다 맛있고 맥주를 부르는맛입니다, 완전맛있어요~~~ㅋㅋ 엽떡즐거먹었는데 이젠 여기서먹을거같아욤!!! 맛있게매운맛 강추, 보통맛도 엽떡처럼 매워요 근데 맛있음 양도많고, 떡볶이 맛있어요. 엽떡만 먹다가 이번에 처음 시켰는데 너무 좋네요 돈가스도 바삭하고..	프랜차이즈 떡볶이의 선두주자를 달리고 있는 '엽기 떡볶이'의 줄임말인 '엽떡'이 떡볶이 가게의 리뷰에서 빈도수가 높음을 확인할 수 있었다. 대부분 '엽떡보다 맛있는 떡볶이'라는 반응이 주를 이루었다.
저녁	도야족발-장안점 도야족발-본점 화룡불닭	출출한 저녁에 먹기엔 강추입니다., 맛있어요 일요일 저녁도 맛나게 끝~~, 수가 맛있네요. 배달시간도 저녁 시간이라 오래걸릴줄 알았는데 빨리 오구요. 깔끔하게..	출출한 저녁에 시켜서 간편하게 혹은 맛있게 먹는다., 저녁인데도 배달이 매우 빨라 서비스가 만족스럽다.는 반응이 주를 이루었다.
스트레스	직화신불닭 화룡불닭 불난떡볶이-광진구점	더운 날씨에 스트레스 확 풀리는 매운맛이네요 맛나요!!!번창하세요, 맛있어용 매워서 스트레스 확~ 날아가용~, 정말 맛있어요~스트레스받을때 맥주랑 팔~ 엄청매워요..매운맛시켰더니ㅋㅋ 맛나게 잘먹었네요. 스트레스 풀려고 먹었는데 굿!	'스트레스'라는 자체의 단어는 부정적인 단어이지만 배달 어플리케이션 리뷰 데이터의 '스트레스'라는 단어는 매운 음식을 먹고 스트레스가 풀렸다.라는 긍정적 의미로 해석됨을 확인할 수 있다.
오시	불난떡볶이-광진구점 도야족발-장안점 미스몽도시락-건대점	와우 25 분안에오시다니!! 굿굿, 눈도 엄청오고 오시기 힘드셨을텐데 받고 맛있게 먹었습니다 감사합니다, 맛있음 비오는날 배달오시느라 수고하셨습니다,	'오시'라는 단어가 문맥적으로 매우 애매한데, 배달 어플리케이션의 리뷰 데이터에서 기상 악화의 경우에도 매우 빠른 배달 속도에 만족하는 긍정적

		배달도 알려주신 시간 맞춰 오시고 맛은 좋았어요	키워드를 의미한다.
이틀	홍닭 직화신불닭 화룡불닭	맛있어서 이틀만에 또시켜먹었어요 양도 진짜많이주시고 친절히 배달해주셔서 정말 맛있게..., 불향이진하고 맛있어서 이삿날 먹곤 이틀만에 다시시켜먹었어요~ 단골해야겠어요., 이틀한번꼴로주문...중독이맞는듯. 매운걸 좋아해서인지 매운맛도 별로안매워용	‘이틀’역시 문맥적으로 긍정적 혹은 부정적으로 분류하기 어려운 단어이지만, 주로 ‘이틀만에 다시 주문하였다.’라는 의미로 긍정적 키워드로 분류되었음을 확인할 수 있다.
베리	HONEY쇼우가족발 피자마스터	베리베리굿, 와우~베리 굿이에요. 진짜맛있다 ㅋ, 베리굿굿 다음날 식어서먹었는데도 맛있어요~, 처음 치고그는 100 점만점에 100 점 드리죠 굿 베리베리 굿, 베리굿이에요 이번이 두번째 시키는건데 음식이 빨리 배달되고 맛도있고 ㅋㅋㅋ	영단어 Very를 소리나는 대로 적은 말로, 한글의 ‘매우’와 유사한 의미를 내포한다. 주로 ‘베리 굿’으로 음식의 맛에 긍정적 평가를 내릴 때 사용되었다.
아저씨	피자마스터 요리왕 화룡불닭	일단 배달하시는 아저씨가 친절하시고~생각보다 배달시간도 빨라서 좋았어요~ 맛은 체인..., 맛도 괜찮고 배달도 빨리오고 배달부아저씨도 친절 하셔서영 ㅌㅌ 앞으로 여기서 시켜먹..	음식을 배달해주시는 기사분들을 ‘배달부 아저씨’라고 칭하며, 배달원 서비스에 대한 긍정적 키워드로 분류되었다.
이용	도야족발-장안점 BHC-중곡대박점 홍닭	양도 많거 배달도 엄청빨라서 자주이용합니당, 가격대비 훌륭한 맛입니다. 자주 이용할께요~, 항상맛있게먹고이용하고있어요 ㅎㅎ 오늘도시켜먹을까생각중입니다	너무 맛있어서 앞으로 자주 이용하겠다는 의지를 밝히거나, 언제나 잘 이용중이라는 문장이 주를 이루었다.
걱정	도야족발-장안점 화룡불닭 불난떡볶이-광진구점	걱정했는데 맛있고 깨끗해요^^, 주문했는데 진짜 맛있고 빠르고 과자에 예쁜글씨까지 감동이네용 좀떨어서 걱정했는데	‘걱정’도 역시 ‘스트레스’와 같이 부정적 단어로 인식되지만, 배달 어플리케이션 내에서 걱정과는

		많이매울까봐 걱정했는데 매콤한 정도네요 맛있어용!!! 맛있었고 빨리오셨고 좋았어요~ 별점 안 좋아서 걱정했는데 다행이네요~기대 이상이었음	달리 매우 맛있다는 의미로, 긍정적 키워드로 분류되었다.
--	--	--	---------------------------------

긍정 키워드 분석의 분석 결과로 보아, 화양동 주민들은 주로 ‘엽떡’과 같이 ‘고추’가 많이 들어간 매운 음식을 먹고 ‘스트레스’를 푸는 행태를 확인할 수 있다. 또한 ‘저녁’에 주문한 음식들에 대해 상대적으로 긍정적인 평가를 내리고 있음을 확인할 수 있다. 맛도 중요하지만 배달의 속도와 배달원의 서비스, 친절함도 중요하게 생각하고 있음을 확인할 수 있었다. 음식을 맛있게 즐긴 고객들은 ‘맛집’, ‘번창하세요’, ‘단골’이라는 단어를 사용하며 가게에 대한 긍정적인 평가를 내렸다. 추가적으로, 굽네치킨-화양점의 매출을 늘리기 위해서 고추바사삭이라는 메뉴의 홍보를 더 극대화하기를 권장한다.

4-2) 부정 키워드 분석

일반적으로 ‘최악’, ‘실망’, ‘별로’, ‘취소’라는 단어는 부정적인 의미로 받아들일 수 있다. ‘그냥’, ‘무난’, ‘괜찮’이라는 단어는 이전 과정에서 단어에 rating을 부여할 때 중립적인 평가를 부정적인 평가로 포괄하여 부정적인 키워드로 분류되었을 것이라고 예상할 수 있다. 그러나 ‘만해’, ‘한지’, ‘예정’, ‘볶음밥’, ‘시간’, ‘그거’, ‘김치찌개’, ‘비빔밥’, ‘불고기’, ‘대비’, ‘단무지’라는 단어를 보면 바로 긍정적인 키워드라고 유추할 수 없다. 해당 단어가 포함되어있는 리뷰 data를 개별적으로 관찰하여 해석하고자 한다.

단어	주 점유 업체	대표 문장	해석
만해	직화신불닭 요리왕 피자마스터	그저그랬어요 한번쯤은 먹어볼만해요, 가격대비 먹을만해요 제 취향의 닭강정맛은 아니었어요,	만해라는 단어는 ‘먹을만해요’라는 단어에서 파생된 것을 확인할 수 있었다. 이 역시 중립적인 평

		먹을만해요 김치찌개 국물이좀 적어요 고기비계많음 그래도 나쁘지 않음	가이므로, 부정적 키워드에 포함되었을 것으로 판단된다.
한지	요리왕 BHC-중곡대박점	주문한지 1 시간 30 분... 아직 도 착안함..., 전화는 통하지도 않고. 주문한지가 두시간 넘었어요. 배달못해주면 문자라도 해야는거 ..., 바쁜건 이해하나... 주문한지 2 시간이 넘었는데... 전화통화도 되지 않고... 이...	‘한지’라는 단어는 주문한지 라는 어구에서 파생되었다. 대체적으로 주문하고 오랫동안 배달을 받지 못해 고객들이 불만을 보임을 확인할 수 있다.
예정	BHC-중곡대박점 짬뽕-통오징어떡볶이 파리에다녀온치킨-건 대점	배달예정시간 50 + 20 분 지각 음식 다 식어서 도착 거기에 김치찌개 포장미숙으로..., 예정시간보다 50 분 더 늦었네요, 다 좋은데 예정시간을 너무 오버합니다 ..., 요기요주문하고 주문 45 분소요예정이라고 안내문자 까지 받았는데 시간이다 지나가고 왜 안...	배달 어플리케이션을 사용하면 예상 배달 도착 시간을 알려준다. ‘예정’이라는 단어 또한 ‘한지’와 같은 의미로 해석될 수 있다.
볶음밥	화양156 요리왕 동강	짬뽕은 역시 굶!!! 볶음밥은 조금 맨밥같았어요πππ, 개최악 쓰레기를 배달해주네요 아직 이런 식당이 있다니 볶음밥 쟁반짜장 시켰는데 한입..., 탕수육 부먹도아닌데 올때부터 엄청 눅눅 짜장보통 볶음밥은 짜장이오래됐는지 신맛남.....	화양동에서 파는 중식집의 ‘볶음밥’에 대한 부정적인 리뷰이다. 짜장면은 보통이나 볶음밥이 최악이다, 볶음밥이 너무 맛없다 라는 반응이 주를 보였다.
시간	요리왕 도야족발-장안점 BHC-자양행복점	맛은 좋은데 배달 시간이 넘 심하네요. 배고프기전에 배달시켜야될 것 같습니다., 맛은있는데 배달 2 시간걸리고.. 상추는 물먹어서 시들시들.. 상추 끝부분 자르고 먹었어..., 한시간다되어가는데 안와요 전화도안받네요...	배달 소요시간에 관련된 단어이다. 배달 시간이 빠르다는 리뷰도 보였지만, 배달 시간이 느리다는 더 리뷰가 많았기에 부정적 키워드로 판단되었음을 확인할 수 있다.
그거	도야족발-장안점 BHC-중곡대박점	피자를 급하게 만들었는지 안잘렸어요ㅋㅋ 그거빠곤 맛있어요~,	‘그거 말곤 괜찮다.’, ‘그거 빠곤 좋았다.’ 라는 문

		너무 늦게 왔어요 한시간 십분? 주소도 못찾구 다 식어서 오구 그 거 만 좀 그랬어요..., 짜장면 면이 불었어요 퓨 그거 말 고는 맛 괜찮아요!	장이 주를 이루었다. 하 나의 단점을 언급하고, 그 점만 제외하면 좋았다는 표현으로 확인되었다.
김치찌개	전주석쇠불고기-본점	맛있어요 양도 많아요 남자 2 명 이면 먹을듯 근데 김치찌개는 제 입맛에는 달아요, 고기많이타서왔어요 김치찌개 너 무셔요 맛은 나쁘진않아요, 김치찌개에 배추대가리가 거의 다 여서 그렇지 ㅋㅋㅋㅋ 맛은 있어 여	해당 데이터에서 김치찌 개를 취급하는 가게는 전 주석쇠불고기 하나였다. 김치찌개도 호불호가 많 이 갈렸는데, 불호의 리 뷰가 더 많아서 부정 키 워드 분류되었다.
불고기	전주석쇠불고기-본점 피자샵-자양성수점	고기가 부위를 다른부위를 사용한 것같이 불고기맛이 나질않더군요, 맛은 있는데... 불고기피자 시켰더 니 야채는 하나도 없이 정말 불고 기만 토핑되어 있..., 그냥 그냥 피자 맛? 불고기는 좀 짜고 달고 그랬어요 토핑도 다 떨 어져서 아쉬웠어요	전주석쇠불고기의 불고기 가 비계가 많다, 맛이 없 다는 표현과 피자샵에서 판매하는 불고기피자의 질이 좋지 않다는 평가가 주를 이루었다.
대비	피자마스터 미스몽도시락-건대점	피자는 가격대비 나쁘지않아요. 스파게티는 별로네요., 가격대비 쏘쏘 토핑이 좀더 많음 좋겠네요, 맛은 보통. 가격 대비 괜찮음, 가격대비 별로..., 무난하지만 가격대비로는 조금 부 족한 것(음료수도 없고)같습니다.	가격 대비 괜찮다, 즉 가 성비가 평범하다. 라는 의미에서 파생된 단어이 다. 이 역시 중립적 평가 는 부정적으로 분류했기 에 생긴 결과로 볼 수 있 다.
단무지	요리왕 불난떡볶이-광진구점	단무지가쉬웠어요 만두는갠차 늬 π, 자장면 겁나짜요. 단무지는 썬거 가지고 오고 깐풍기는 닭껍질만 있네요. 아무튼 별로..., 진짜 단무지좀 많이달라니까 그게 아깝나 어이가없다 진짜	자장면, 떡볶이와 함께 배달되는 단무지의 질에 대한 비판적인 평가로 인 하여 부정적 키워드로 분 류되었다.

부정 키워드 분석의 분석 결과로 보아, 화양동 주민들은 ‘한지’와 ‘예정’, ‘시간’의

해석 결과로 ‘배달 서비스’에 불만을 강하게 가짐을 확인할 수 있었다. 긍정적 키워드 분석에서는 ‘부정적 단어’로 해석될 수 있는 단어가 등장했지만 부정적 키워드 분석에서는 ‘긍정적 단어’로 해석될 수 있는 단어는 찾아보기 어려웠다. 또한, ‘김치찌개’, ‘불고기’, ‘단무지’와 같이 특정 가게에서 판매하는 특정 메뉴에 국한된 리뷰가 부정 키워드의 계수가 높게 나타난 것을 확인할 수 있었다. 전주석쇠불고기는 김치찌개와 불고기의 맛에 더욱이 신경을 쓸 필요가 있어 보인다. 요리왕과 불난떡볶이에서는 단무지의 품질에 마찬가지로 주의를 더 기울여야 할 것이다.

V. 결론 및 한계점

본 과제에서는 소비자들이 직접 음식을 먹고 느낀 감정들을 리뷰 코멘트와 세부 항목별(맛, 양, 배달) 점수에 담아 가게의 음식 및 서비스에 대한 평가한 데이터의 감성을 분석하여 소비자 만족도 제고 방향성을 제시하고자 하였다. 리뷰 코멘트가 존재하지 않는 데이터도 있었고, 맛, 양, 배달 점수가 존재하지 않는 경우도 있었다. 점수만으로 리뷰의 코멘트를 예측할 수 없다고 판단하였기에 리뷰 코멘트가 미존하는 데이터는 제거하고, 리뷰에 특정 단어가 들어가지만 항목별 평가 점수가 존재하지 않는 데이터는 특정 단어가 들어갔고, 항목별 평가 점수가 존재하는 데이터의 산술평균으로 점수를 대체하여 감성분석 모델을 제작하였다. 이외에도 필터링되지 못한 데이터는 '□□□□□'와 같이 감성을 예측할 수 없거나, 특수문자나 중국어 또는 외계어로 적힌 리뷰가 존재하였다. 해당 데이터는 제거를 진행했다. 빈도 분석을 통해 배달 어플리케이션 사용자 리뷰에 포함된 주요 요인을 추출하였고, TF-IDF 기법을 이용하여 주요 요인에 대한 리뷰 내 상대적 중요도를 가중치로 추출하였다. 긍정적인 리뷰가 부정적인 리뷰보다 데이터의 개수가 매우 많았기에, 부정적인 리뷰도 긍정적인 리뷰로 예측하는 모델이 형성되었다. 따라서 1:1 Sampling 기법을 사용하였는데, Confusion Matrix를 사용하여 예측한 결과는 절반보다 조금 높은 정도의 평가 척도를 보였다. 이후, 로지스틱 회귀분석을 사용하여 맛, 양, 배달의 산술평균이 4점 이상인 리뷰들은 긍정적 리뷰로 분류하였고, 그렇지 않은 리뷰들은 부정적 리뷰로 분류하였다.

전체적으로 화양동 배달 어플리케이션의 키워드 분석의 결과를 조합해보자면 양과 가격에 대한 평가는 맛과 배달에 대한 평가에 비해 현저히 적은 리뷰를 보였다. 맛이 있어도 배달 소요시간이 오래 걸릴수록 부정적인 평가를 내리는 추세를 확인할 수 있었으며, 즐겁게 먹은 음식은 가격과 관련된 리뷰를 거의 찾아볼 수 없었지만 불만족스럽게 즐긴 음식에는 가격과 관련된 리뷰를 찾아볼 수 있었다. 긍정적 키워드와 부정적 키워드를 4개의 카테고리(맛, 배달, 가격, 서비스)로 분류하면 다음과 같이 정리할 수 있다. 가게 업주들은 메인 메뉴의 맛에만 치중하지 않고 단무지, 치킨 무, 소스와 같은 부속적인 음식에도 신경을 써야할 것이다. 추가로 배달의 속

도도 빠른 배달을 위해 다른 방안을 고려하거나, 느린 배달 시스템을 개선한다면 화양동 소비자의 만족도를 제고할 수 있을 것이다.

	긍정적 키워드	부정적 키워드
맛	대박, 고추, 엽떡, 맛집, 저녁, 스트레스, 이틀, 베리, 치즈, 날치, 홍닭, 짬뽕, 불닭, 야들야들, 걱정	최악, 그냥, 만해, 실망, 보통, 무난, 볶음밥, 별로, 그거, 김치찌개, 비빔밥, 괜춘, 불고기, 단무지
배달	저녁, 오시, 아저씨,	한지, 예정, 시간, 그거, 취소
가격		대비
서비스	변창, 단골, 이용	최악, 실망, 별로, 그거

그러나, 감성분석에서 본 과제는 보고 결과를 해석하는 데 있어 다음과 같은 한계점을 갖는다.

첫 번째, 본 보고는 배달 어플리케이션 사용자 리뷰를 이용한 연구로 중요한 요인을 빈도 분석과 감성분석을 통해 파악하여 실무자에게 해결책을 제안하고 있다. 사용자 리뷰는 소비자 관점에서 작성된 것으로 모든 관점을 아우를 수 없다는 한계가 있다. 이에 소상공인 관점에서 제기되는 문제나 환경적, 사회적으로 다뤄질 수 있는 문제점을 알아보기 위한 연구가 필요하다.

두 번째, 온라인 리뷰에 대해 빈도 분석을 이용하여 주요 요인을 추출하였다. 빈도 분석으로 추출한 키워드는 단순한 빈도에 기반한 것으로 한 문장 내 여러 키워드가 존재하는 경우에 대해서 고려하지 않는다. 즉, 여러 키워드를 포함하는 하나의 문장을 빈도 분석했을 때, 하나의 리뷰가 여러 키워드를 대표하는 경우가 발생한다. 정확한 분석을 위해 리뷰 내 단어의 가중치를 산출하고, 가중치가 높은 단어가 해당 리뷰를 대표하며 다른 키워드를 통해 중복적으로 리뷰가 다뤄지지 않도록 연구가 필요하다.

세 번째, 주요 요인의 상대적 중요도와 감성분석의 감성 극성 값을 결합하여 산출한 새로운 데이터는 오차가 발생할 수 있다. 예를 들어 “배달이 빨라서 너무 좋은데, 수수료가 비싸고 직원 태도가 별로예요.”와 같은 리뷰가 존재하고 감성 극성 값이 -0.6 이라고 가정했을 때, 감성 극성 값과 주요 요인의 상대적 중요도를 결합하면 배달 키워드에 대해 긍정적인 반응이었음에도 새로 산출된 데이터에는

배달 키워드에 부정적인 반응을 나타낸 것으로 나타나게 된다. 즉, 실제로는 긍정적인 키워드가 새로 산출되는 데이터에서는 부정적인 키워드로 인식될 수 있다. 이러한 오차를 없애기 위해 한국어의 단어 레벨의 감성 상태를 추출하기 위한 연구와 한국어 감성어 사전 구축이 필요하며, 구축된 사전을 통해 높은 정확도의 연구가 필요하다.

네 번째, 리뷰 이벤트로 인해 점수에 왜곡이 있을 가능성이 존재한다. 처음 보고를 시작할 때에도 알 수 있듯 부정적인 평가보다 긍정적인 평가가 매우 많이 존재하였다. 애초에 긍정적인 평가 위주로 이루어진 데이터이므로 Confusion Matrix 에서도 좋지 못한 성능을 보였다. 이는 가게 및 업체들의 맛집 랭킹 상승을 위해 진행하는 리뷰이벤트의 가능성을 배제하지 못한다.

다섯 번째, 형태소 및 명사가 제대로 분리되지 못한 값들이 존재한다. 감성사전은 말그대로 사전적인 의미를 가지는 단어들로 구성되어 있을 것이다. 그러나, ‘주문한지 2시간이 되어가는데...’라는 문장에서 ‘주문’, ‘한지’, ‘시간’으로 나뉘어지는 행태를 보았을 때, 결코 단어들이 제대로 분리되었다고 단언할 수 없다. 감성사전은 ‘한지’라는 단어를 위와 같은 의미로 받아들이지 않고, 동음이의어 한지, 즉 한국의 종이라는 뜻을 가진 韓紙로 해석했을 가능성도 존재한다. 이는 세 번째 한계점과 연결된다. 이러한 오차를 줄이기 위해 발전된 연구가 필요하다고 여겨진다.

References

- [1] 김장혁, 박수진, & 김철민. (2020). 감성 분석 모델을 적용한 숙박 애플리케이션 리뷰 분석 서비스. *한국컴퓨터교육학회 학술발표대회논문집*, 24(2 (A)), 123-126.
- [2] So, J. S., & Shin, P. S. (2020). Rating prediction by evaluation item through sentiment analysis of restaurant review. *Journal of the Korea Society of Computer and Information*, 25(6), 81-89.
- [3] Hong, T. (2022). Sentiment Analysis and Star Rating Prediction Based on Big Data Analysis of Online Reviews of Foreign Tourists Visiting Korea. *Knowledge Management Research*, 23(1), 187-201.
- [4] 신동헌. (2017). TF-IDF 기반 이직 대상 기업 추천 시스템 설계 및 구현: 잡플래닛 리뷰를 중심으로
- [5] 이윤주, 김희진, 이예슬, & 정혜선. (2021). 로지스틱 회귀모형과 의사결정 나무모형을 활용한 청소년 자살 시도 예측모형 비교: 2019 청소년 건강행태 온라인조사를 이용한 2 차 자료분석. *Journal of Korean Academy of Nursing*, 51(1), 40-53.