아이디어 제안서

제안자	강미나 (UNIST, kasong13@unist.ac.kr), 조민수 (UNIST, minsu3123@unist.ac.kr), 권기범 (UNIST, kwonrjatk@unist.ac.kr), 허보인 (UNIST, qhdls5340@unist.ac.kr), 고지현 (UNIST, kjh1337@unist.ac.kr), 이승환 (UNIST, tmdghks3155@unist.ac.kr)
아이디어 제안 명	LG 전자 Growth Hacking 실시간 모니터
제품 유형	소프트웨어
I. 제안 배경	I.1 제품의 필요성 (개발목적) 및 현황 소비자들은 구매에 관한 의사결정에서 고객의 제품 리뷰를 중요한 평가 지표로 활용한다. 쉽고 빠르게 본인의 의사를 표현할 수 있다는 장점으로 인해 온라인 제품 리뷰 데이터베이스는 점점 방대해지고, 마케팅 및 제품 개선에서 이러한 제품 리뷰의 확보 및 활용의 중요도는 날로 커지고 있다. 그로스 해킹 (Growth Hacking)은 데이터에 기반하여 소비자와 가장 밀접한 문제를 파악하고 해결해 고객의 요구를 충족시키는 마케팅 기법이다. 다양한 종류의 데이터가 그로스 해킹에 사용되며, 주로 서비스와 제품에 대한 소비자 기반 데이터가 분석에 사용된다. 대중적인 방법으로는 쇼핑몰 등의 인터넷 쇼핑 플랫폼에서의 사용자 평가 데이터를 기반으로 하는 소비자 분석 방법이 알려져 있다. 이를 통해 빠르게 변화하는 시장의 흐름과 소비자의 요구를 파악하고, 이에 대응할 계획을 빠르게 수립하는 데에 있어 논리적 기반이 될 수 있다. 예를 들어, 인터넷 쇼핑 플랫폼 제품 리뷰 데이터에서 추출된 토픽의 파악은 기업이 제품 개선과 판매율 증가를 위하여 어떤 전략이 필요한지를 파악하는 데에 중요한 단초를 제공할 수 있다.  Figure 1. 인테리어 쇼핑 플랫폼에서의 리뷰 사례 불과 2010년대 초반까지는, 제품 리뷰를 바탕으로 한 기업의 제품 개선은 오로지 연구자들의 수작업으로 이루어졌기에 시간적·비용적 한계가 있었다. 뿐만 아니라, 리뷰를 분석하는 데에 있어 연구자의 주관적인 성향이 개입되는 경우도 존재한다. 이를 방지하기 위해 제안된 제품 아이디어에는 기계학습을 이용한 리뷰 토픽 모델링 기법을 사용하였다.

I. 제안 배경

제품의 리뷰들을 바탕으로 개선점과 문제점을 인식하고 부정적인 반응이 많이 나오는 토픽을 추려낼 수 있어야 한다. 위에서 제시한 토픽 모델링 기법을 통해 얻은 결과를 바탕으로 감성 분석을 실시하여 한 묶음의 토픽에 대하여 긍정도와 부정도를 측정할 수 있고, 이는 제품 개선 방안 우선순위 결정의 근거가 될 수 있다.

따라서 제안된 제품 아이디어를 이용하면 분석가는 다양한 플랫폼에 게시된 많은 수의 리뷰를 정량·정성적이고 객관적인 방법으로 단시간에 분석할 수 있으며, 이를 통해 소비자가 게시한 제품 리뷰에 내재되어 있는 잠재적인 의미를 추론할 수 있다. 이 과정을 자동화하여 실시간으로 변화하는 소비자들의 긍정·부정 반응 및 소비자 요구 경향을 빠르게 파악할 수 있다.

이 프로젝트는 LG 전자의 퓨리케어 제품인 '에어로타워 (2021.12)'를 대상으로 진행되었으며, 데이터 수집부터 분석 및 시각화의 단계를 모두 GUI 상에 자동화하여, 검색 키워드만 바꾸어 입력하면 다른 제품에 대해서도 같은 방식으로 분석할 수 있다는 편리함이 있다.

I.2 관련 연구 및 적용 사례

토픽 모델링을 이용한 리뷰 마이닝은 이미 다양한 연구에서 진행되어왔다. '텍스트 마이닝 기법과 ARIMA 모형을 활용한 배달의 민족 앱 리뷰 분석' 연구에서는 다음과 같은 단계로 리뷰 분석을 진행하였다. 첫 번째로 데이터 수집, 두 번째로 데이터 전처리 (불용어 제거, 형태소 분석), 세 번째로, LDA 토픽 모델링, 마지막으로 시계열 이상치 탐지, 감성 분석 등이 진행되었다. 데이터로는 구글 플레이 스토어 및 애플 앱 스토어에 게시된 배달의 민족 앱 리뷰를 이용하였다. 데이터 수집 후 전처리 과정에서 정규 표현식을 이용하여 같은 단어가 반복되는 문장을 정제하였고, 5개 이상의 명사를 갖는 리뷰를 분석을 위한 데이터로 이용하였다.

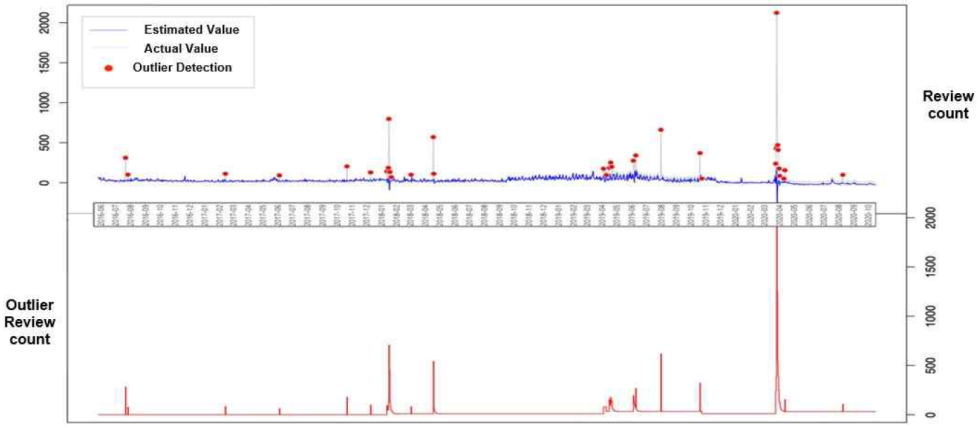


Figure 2. 시계열 이상치 탐지 결과

데이터 분석 과정에서 전체 리뷰의 주요 토픽을 LDA를 이용하여 정제하였다. 이후 시계열 이상치 탐지를 통해 리뷰 수가 많은 날짜를 탐지하여 해당일의 감성 분석을 진행하였고 이를 일자 별 발생 이슈와 비교하였다. 그 결과, 기업 윤리나 음식점 서비스 관련 토픽에서 감성 분석 수치가 부정적인 경우가 많았고 연구팀은 이에 대한 개선책을 제시하였다.

Ⅱ. 아이디어 제안

Ⅱ.1 개발 내용 및 범위

a. 데이터 수집

LG 전자 에어로타워의 리뷰 데이터를 얻기 위해 처음으로 접근한 방식은 LG 전자 에어로타워 관련 유튜브 영상 댓글을 수집하는 것이었다. 하지만, 유튜브 댓글은 데이터의 양이 많지만, 의미가 없거나 사용자 리뷰로써 가치가 없는 데이터들이 많았고, 논리성・명료성이 결여된 구어체를 사용한 평가가 많았기에, 인터넷 쇼핑 플랫폼의 리뷰 데이터로 데이터 수집 방향을 변경하였다. 인터넷 쇼핑 플랫폼의 리뷰는 데이터의 양은 적지만, 실사용자가 직접 작성한 의미 있는 데이터였기에 향후 분석을 할 때 더욱 도움이 될 것으로 판단하여 인터넷 쇼핑 플랫폼에서 리뷰 데이터를 수집하였다. LG 전자 에어로타워 실사용자의 리뷰 데이터를 수집하기 위해 해당 제품을 판매 중인 인터넷 쇼핑 플랫폼에서 리뷰 데이터에 대한 완전 자동화된 크롤링을 진행하였다. 예시로 선정된 플랫폼인 네이버 쇼핑, 하이마트, 오늘의 집, 다나와에서 리뷰 데이터를 수집했다. 크롤링할 데이터는 [date, reviews, star] 양식으로 수집하였다. 데이터 크롤링의 방식은 쇼핑몰마다 약간의 차이가 있는데, 네이버 쇼핑, 오늘의 집, 다나와를 대상으로는 공통적으로 Selenium과 BeautifulSoup Python 라이브러리를 기반으로 개발하였다. Selenium의 Chromedriver를 통해 가상의 환경을 설정하고 스크롤, 페이지 넘기는 과정을 반복하며 모든 데이터를 수집했다. 하이마트의 경우는 최종 결과물인 GUI 상에 완전한 백엔드(back-end) 형태의 결합을 위해 데이터 수집 과정에서 Selenium의 Chromedriver를 사용하지 않고 Get・Post Request를 통한 웹페이지 송수신 데이터를 직접 분석하여 크롤링을 진행하였다. LG전자 에어로 타워의 출시일인 2021년 12월 23일부터 2022년 3월 23일까지의 리뷰 데이터 기준으로 크롤링을 통해 확보한 데이터는 네이버 쇼핑 270개, 하이마트 51개, 오늘의 집 61개, 다나와 109개로 총 491개를 확보하였다.

	date	review	star
2	20220316	배송이 걸리긴했지만 선물받으신 분이 만족한다고 합니다	5
3	20220314	좋아요~ 필터 여유 하나 더 있었음 좋을듯ㅎㅎ	5
4	20220310	디자인 이쁘고 조용하고 정말 좋아요~추울때 샀으면 온풍기능을 잘썼을꺼 같은데 오늘부터 날이 따뜻해서 다음 겨울에나 쓰겠네요 ㅎㅎ	5
5	20220310	핫딜에 저렴히 구매 했습니다ㅎㅎ감사합니다.진작에 살 것을ㅜㅜ	5
6	20220310	잘산거같아요 더시용해봐야알겠지만	5
7	20220310	첫 째방이 쫌 추운편이라 꽁꽁 싸매고 재우다가 도저히 안되겠다 싶어서 찾아보니 공기청정기도 되고 온풍도 사용할 수 있는 에어로타워기	5
8	20220310	디자인 성능 두루두루 맘에 들어요~~	5
9	20220310	크기도 적당하고 소음도 적어요	5
10	20220309	배송 빠르고 설치 10분도 안걸렸어요~온풍 기능 너무 좋으네요디자인은 말할 것도 없어 인테리어에 최고입니다.리모컨 너무 앙증맞고 부	5
11	20220307	배송도 엄청 빠르고요아들이 좋다고 합니다.디자인이 넘 예쁘다네요.	5
12	20220307	저희 아파트 아트월과 찰떡입니다 고민은 배송만 늦출뿐이었습니다. 넘 맘에 드네요! 배송 직원분도 친절하셨습니다^^	5
13	20220306	배송 빠르고 제품이 좋네요	5
14	20220305	만족합니다 따뜻한바람 좋아요	5
15	20220304	좋아요 따뜻한 온풍이 금방 전달되요	5

Figure 3. 하이마트 플랫폼의 데이터 수집 예시

b. 데이터 라벨링

감성 분석 (Sentimental Analysis) 기법을 위한 특징 변수를 소비자의 긍정・부정 반응으로 선정하였고, 이러한 특징 변수를 추론하기 위해 데이터 수집 과정에서 별점을 수집하였다. 하지만, 수집한 데이터의 별점의 분포가 만점 (5점)에 가깝게 상향평준화 되어있었고, 아쉬운 점이나 불만 사항이 있어도 대부분 3점에서 4점의 별점이 부여되어있었다. 또한, 한 문단 안에 긍정과 부정의 의미를 가진 문장이 함께 있는 경우도 있었다. 이를 해결하기 위해 문단 단위의 평가를 문장 단위의 평가로 토큰화를 진행하며 제안자가 직접 긍정적 문장을 1로, 부정적 문장을 -1로 이진화 하였다. 이를 통해 491개의 문단 데이터를 1431개의 평가된 문장으로 분리하였다.

	date	review	rate
2	20220224	선물로 줬는데 사용해보니 정말 만족스럽다고 하네요.	1
3	20220224	추천합니다	1
4	20220222	배송도 빠르고	1
5	20220222	기사님도 친절하셨어요.	1
6	20220222	디자인 너무너무 맘에들구요.	1
7	20220222	엘지 좋아요	1
8	20220217	오후에 주문하고 다음날 3시에 배송되고 설치완료됨.	1
9	20220217	기사님이 다 설명해주시고 필터관련 설명들음.	1
10	20220217	매우만족	1
11	20220216	인기가 무지막지해서	1
12	20220216	주문 후 한달은 걸렸어요.	-1

Figure 4. 이진화 및 문장별로 나뉘어진 데이터

c. 데이터 전처리

데이터 전처리 과정에서는 데이터 분석 방식에 적합한 형태로 데이터를 정제하는 것을 목표로 하였다. 데이터 분석 방식이 감성 분석 및 토픽 모델링이므로, 이에 걸맞게 데이터를 수정하였다.

형태소 분석이란, 형태소를 비롯하여 어근, 접두사·접미사, 품사 등 다양한 언어적 속성으로 문장의 구조를 파악하는 것을 의미한다. 문장 분석에서 불필요한 조사, 부사, 감탄사 등을 모든 문장에서 제거하고 형태소 분석을 진행했다. 형태소 분석은 KoNLPy 기반의 Mecab 한글 분석기를 사용하여 진행하였다. Mecab 분석기는 C 기반 연산을 바탕으로 KoNLPy에 포팅된 엔진으로, KoNLPy 기반의 다른 형태소 분석기 (Hannanum, Kkma, Twitter 등)와 비교했을 때 수행 시간이 탁월하게 빠르다는 특징이 있다. 특히, 형태소 분석기의 경우에는 단어의 개수가 늘어날수록 수행 시간이 기하급수적으로 증가하는 경향이 있는데, Mecab 분석기의 경우에는 단어의 개수가 많아진다 하더라도 비교적 빠르게 형태소 분석을 수행한다.

II. 아이디어 제안

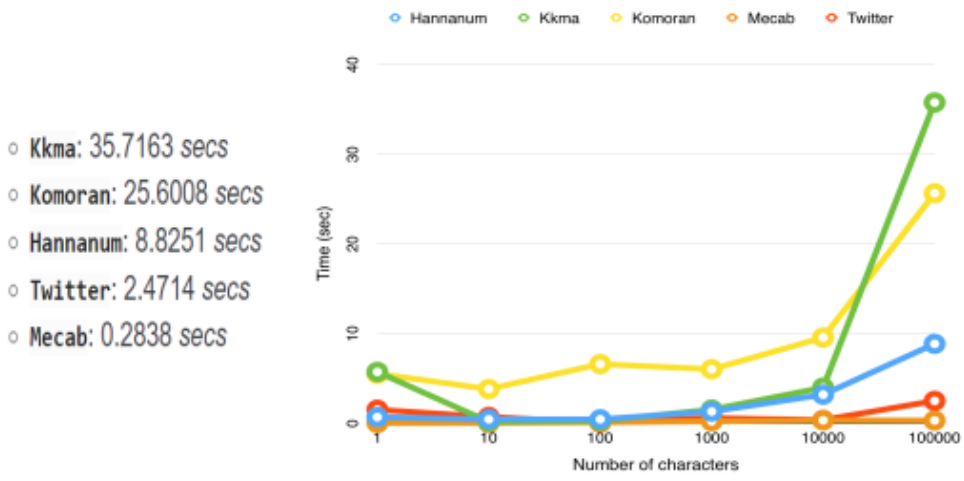


Figure 5. 형태소 분석기별 수행 시간

Mecab 분석기는 수행 시간이 빠를 뿐만 아니라 형태소 분석의 성능 또한 준수하다. 다음은 “아버지가방에들어가신다” 라는 예시 문장을 형태소 분석기별 분석 결과를 비교한 것이다.

Hannanum	Kkma	Komoran	Mecab	Twitter
아버지가방에 들어가 / N	아버지 / NNG	아버지가방에 들어가신다 / NNP	아버지 / NNG	아버지 / Noun
이 / J	가방 / NNG		가 / JKS	가방 / Noun
시ㄴ다 / E	에 / JKM		방 / NNG	에 / Josa
	들어가 / VV		에 / JKB	들어가신 / Verb
	시 / EPH		들어가 / VV	다 / Eomi
	ㄴ다 / EFN		신다 / EP+EC	

Figure 6. 형태소 분석기별 성능 비교

다른 형태소 분석기와 비교했을 때, Mecab 분석기의 형태소 분석 성능이 가장 준수한 것을 확인할 수 있다. Mecab 분석기는 품사 태깅을 통해 각 단어 옆에 NNG, JKS, VV 등의 품사를 출력한다. 각 품사 태그의 의미는 Table 1.을 통해 확인할 수 있다.

대분류	세종 품사 태그		mecab-ko-dic 품사 태그	
	태그	설명	태그	설명
체언	NNG	일반 명사	NNG	일반 명사
	NNP	고유 명사	NNP	고유 명사
	NNB	의존 명사	NNB	의존 명사
			NNBC	단위를 나타내는 명사
	NR	수사	NR	수사
	NP	대명사	NP	대명사
용언	VV	동사	VV	동사
	VA	형용사	VA	형용사
	VX	보조 용언	VX	보조 용언
	VCP	긍정 지정사	VCP	긍정 지정사
	VCN	부정 지정사	VCN	부정 지정사
관형사	MM	관형사	MM	관형사
부사	MAG	일반 부사	MAG	일반 부사
	MAJ	접속 부사	MAJ	접속 부사
감탄사	IC	감탄사	IC	감탄사
조사	JKS	주격 조사	JKS	주격 조사
	JKC	보격 조사	JKC	보격 조사
	JKG	관형격 조사	JKG	관형격 조사
	JKO	목적격 조사	JKO	목적격 조사
	JKB	부사격 조사	JKB	부사격 조사
	JKV	호격 조사	JKV	호격 조사
	JKQ	인용격 조사	JKQ	인용격 조사
	JX	보조사	JX	보조사

Table 1. Mecab 분석기의 품사 태그

Mecab 분석기를 사용하여 형태소 분석을 한 뒤, 데이터 전처리를 크게 두 가지 경우로 나누어 진행하였다. 첫 번째는, 토픽 모델링을 통한 토픽 추출에 적합한 데이터이고, 두 번째는 감성 분석에 적합한 데이터다. 토픽 모델링을 하는 과정에서 동사와 형용사 등의 용언은 불필요하기 때문에 토픽 모델링을 위한 데이터는 형태소 분석이 완료된 전체 리뷰 데이터에서 일반 명사 (NNG)와 고유 명사 (NNP)만 추출하여 데이터를 정제하였다. 다음으로, 감성 분석의 경우에는 용언을 통해 문장의 의미를 파악할 수 있으므로, 일반 명사 (NNG), 고유 명사 (NNP), 동사 (VV), 형용사 (VA)를 추출하여 데이터를 구성하였다.

Ⅱ. 아이디어 제안

	date	review	rate
0	2022-02-24 00:00:00	선물로 줬는데 사용해보니 정말 만족스럽다고 하네요.	1
1	2022-02-24 00:00:00	추천합니다	1
2	2022-02-22 00:00:00	배송도 빠르고	1
3	2022-02-22 00:00:00	기사님도 친절하셨어요.	1
4	2022-02-22 00:00:00	디자인 너무너무 맘에 들고요.	1
...	...	...	...
1426	2022-01-18 00:00:00	디자인도 공기청정기 중 가장 이쁜 거 같아요	1
1427	2022-01-18 00:00:00	디자인이 너무 예뻐요	1
1428	2022-01-18 00:00:00	선물시켰는데 받는 이가 좋아하네요	1
1429	2022-01-17 00:00:00	배송이 늦었지만	-1
1430	2022-01-17 00:00:00	아주 고급스럽고 작동도 만족합니다.	1

	review	rate
0	[선물, 사용, 만족]	1
1	[추천]	1
2	[배송]	1
3	[기사, 친절]	1
4	[디자인, 맘]	1
...	...	...
1426	[디자인, 공기, 청정기]	1
1427	[디자인]	1
1428	[선물]	1
1429	[배송]	-1
1430	[고급, 작동, 만족]	1

Figure 7. a) 토픽 모델링용 형태소 b) 감성 분석용 형태소

형태소 분석을 진행한 뒤에 명사만 추출하였음에도 불필요한 단어가 남아있었다. 모델링 과정에서 불필요한 단어는 성능에 부정적인 영향을 줄 수 있기에 이러한 단어를 제거하였다. 분석을 하는 것에 있어서 큰 도움이 되지 않거나 기여하는 바가 없는 단어를 불용어 (Stopwords)라고 하는데, 불용어는 직접 불용어 사전 (dictionary)을 만든 뒤 제거하여야 한다. 아래의 Figure 8.은 제안자가 직접 작성한 불용어 사전이다. 예를들어, '신혼'이라는 단어는 LG 전자 에어로타워가 신혼집에 인기가 많다는 등의 유의미한 분석을 가능하게 하지만, '남편'이라는 단어는 분석에 큰 도움이 되지 않으므로 제거하였다. 이러한 방식으로, 모델링에 직접적인 영향을 주는 단어가 아니라면 제거를 하였다.

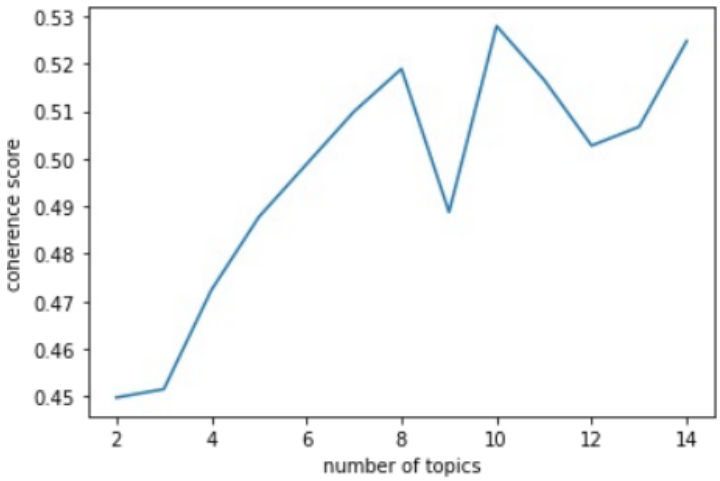
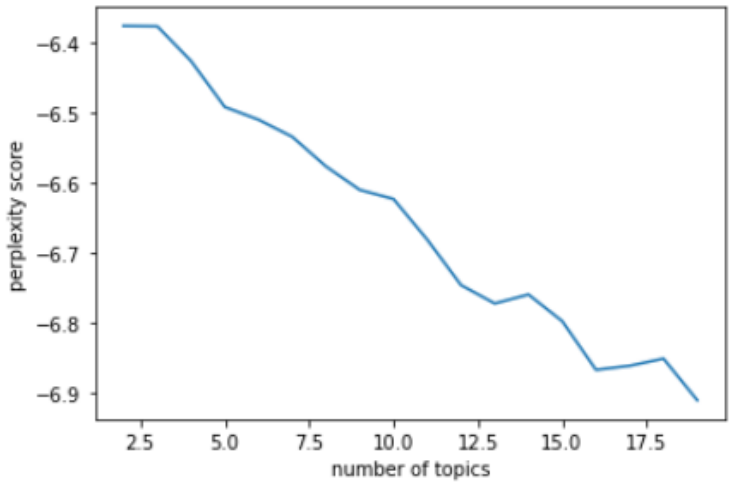
```
stopwords = ['약간', '최근', '딸', '밍', '부', '때', '후', '달', '친언니', '전', '부탁', '아들', '나중', '덕분', '위',
       '대신', '조카', '상의', '남편', '플로', '면', '여기저기', '오늘', '하이', '플', '점', '큐', '감', '이곳저곳',
       '돌', '몰', '앞', '비람', '씽', '훈기', '행', '이후', '기', '셔', '두말', '입', '애', '남', '거사', '이해', '전달',
       '팬', '기', '다음', '날', '배', '면', '플', '뒤', '마왕', '가요', '신념', '속', '톳', '이번', '기회', '신', '차',
       '선짝', '안', '눈', '월', '뭔', '밑', '동생', '제주도', '정도', '옆', '곳', '예', '달', '히', '도', '기와',
       '막', '댁', '사', '오른쪽', '왼쪽', '상도', '요번', '티브이', '그간', '키', '철', '라모', '에어로타워컨', '청', '에어로타워', '타워']
```

Figure 8. 불용어 사전

d. 잠재 디리클레 할당 (Latent Dirichlet Allocation, LDA)을 활용한 토픽 모델링

토픽 모델링은 문서의 집합에서 토픽을 찾아내는 프로세스를 말한다. LDA는 토픽 모델링의 대표적인 알고리즘이다. 이 알고리즘은 단어가 특정 토픽에 존재할 확률과 문서에 특정 토픽이 존재할 확률을 결합 확률로 추정하여 토픽을 추출한다. 즉, 주어진 문서에 대하여 각 문서에 어떤 주제들이 존재하는가에 대한 확률 모델이라고 할 수 있다. 이 알고리즘을 사용하는 데에 있어 분석가가 고려해야 하는 하이퍼파라미터 (Hyper-Parameter)로는 최적의 토픽 수가 있다.

최적의 토픽 수를 결정하기 위해서는 두 가지 정량적인 지표를 보며 판단해야한다. 두 가지 지표로써 Coherence와 Perplexity가 존재하며, 이들을 모두 고려하여 정확도를 높일 수 있는 토픽의 수를 결정한다. Coherence 값에 의존하여 토픽 수를 결정하게 되었을 경우에는 문서의 주제를 반영하지 못하거나, 편향이 나타날 수 있기 때문에 Perplexity를 동시에 고려하여 최적의 토픽 수를 결정하는 과정을 거쳤다. 또한, 토픽 수에 따라 성능을 극대화 할 수 있는 파라미터를 찾기 위해 파라미터 튜닝을 동시에 진행하여, 최적의 토픽 수와 그 수에 맞는 파라미터 값을 찾는 과정을 거쳤다.

Ⅱ. 아이디어 제안	e. Coherence Coherence는 주제의 일관성을 의미한다. n개의 토픽을 설정하였을 때, 해당 토픽 내에서의 상위 단어 간의 유사도를 나타내는 지표이다. 따라서 해당 토픽 내에서 의미론적으로 일치하는 단어끼리 모여 있는지 알 수 있는 지표이다. Coherence 값은 높을수록 토픽 내의 단어 간 의미론적 일관성이 높다고 할 수 있다. 따라서, LDA 모델이 얼마나 실제로 의미 있는 결과를 나타냈는지에 대한 객관적인 지표로 활용할 수 있다. 하지만 문서 집합의 Coherence가 높아지면 단조로운 결과로 수렴하는 문제점이 생긴다. Coherence가 너무 높아지면 정보의 양이 줄어들게 되고, Coherence가 너무 낮으면 정보들이 연관성이 낮아지기 때문에 적절하게 값을 선정해야한다. 다음 그림은 토픽 수에 따른 Coherence 값을 나타낸 그래프이다. 그래프를 통하여 토픽의 수가 약 3~4개일 때 가장 높은 Coherence를 갖는다는 것을 확인할 수 있다.  Figure 9. 모델의 토픽 수-Coherence 그래프 f. Perplexity Perplexity는 혼란도를 의미한다. LDA 모델이 실제로 관측되는 값을 얼마나 잘 예측하는지를 뜻한다. 즉, 주어진 문서의 토픽을 얼마나 잘 반영하는지에 대한 객관적 지표이다. Perplexity 값이 작으면 토픽 모델이 문서를 잘 반영된다고 할 수 있다.  Figure 10. 모델의 토픽 수-Perplexity 그래프

g. Parameter Tunning

Coherence가 크면 클수록 토픽 내의 단어 유사도가 높고, Perplexity 값이 작을수록 실제 문서 내의 토픽을 잘 반영할 수 있다. 하지만, 한계 Perplexity가 낮다고 해서, 결과의 해석이 용이하다는 의미가 아니기 때문에, 다양한 파라미터와 평가 지표를 바탕으로 토픽을 결정하여야 한다. 이에 Heuristic Method를 이용하여 Coherence 값과 Perplexity 값을 동시에 수렴시키는 최적의 파라미터를 찾는 과정을 거쳤다. 두 가지 파라미터의 조합에 따라 LDA 모델의 성능이 결정되기 때문에, 가능한 조합을 모두 시도해 보았다. Alpha는 문서 토픽 밀도를 나타내는 사전 집중 파라미터이다. Alpha 값이 클수록 문서는 더 많은 토픽으로 구성되어 문서 당 더 구체적인 토픽 분포를 얻을 수 있다고 가정한다. Beta 파라미터는 Alpha와 비슷한 개념으로 토픽 워드의 밀도를 나타내는 파라미터이다. Beta 값이 높으면 토픽은 대부분의 워드로 구성되어 토픽 당 보다 구체적인 워드의 분포를 얻을 수 있다. 토픽의 수, Validation_Set, Alpha 파라미터, Beta 파라미터 들의 조합을 통하여 Coherence의 값을 가장 크게 만드는 파라미터와 토픽 수를 결정하였다. 총 540개의 가능한 조합들 중 Coherence가 높으면서, 실제 LDA 모델을 실행하였을 때 토픽들이 유의미한 결과를 가지며, 단어들이 서로 겹치지 않도록 토픽 수와 파라미터들을 조절하였다. Table 2.는 파라미터의 모든 조합 중에서 Coherence가 높은 상위 8개의 조합 결과를 나타낸다. Trade-off를 고려하여 최종 선정된 파라미터 값의 조합은 [num_topics=4, iterations=750, alpha=0.31, eta=0.01]이다. 이 조합이 가장 토픽과 단어의 관련성이 높으며 가장 유의미한 결과라고 판단하였다.

II. 아이디어 제안

Validation_Set	Topics	Alpha	Beta	Coherence
75% Corpus	4	0.01	0.01	0.517451
75% Corpus	4	0.01	0.31	0.524128
75% Corpus	4	0.01	0.61	0.509111
75% Corpus	4	0.01	0.91	0.509188
75% Corpus	4	0.01	symmetric	0.519196
75% Corpus	4	0.31	0.01	0.566448
75% Corpus	4	0.31	0.31	0.535742
75% Corpus	4	0.31	0.61	0.539129

Table 2. 각 조합에 따른 Coherence 결과

h. 토픽 추출

앞에서 얻은 최적의 파라미터 조합을 모델의 파라미터로 적용하여 LDA 모델을 학습하였다. 이 모델을 바탕으로 4개의 토픽을 추출하였다.

	topic_0	topic_1	topic_2	topic_3
0	온풍	디자인	배송	만족
1	사용	공기	구매	기능
2	기능	온풍	친절	온풍
3	디자인	기능	기사	제품
4	제품	청정기	설치	생각
5	배송	청정	감사	가격
6	맘	에어로타워	공기	선물
7	집	타워	디자인	감사
8	소음	만족	청정기	설치
9	생각	사용	집	사용

Figure 11. LDA 모델에서 추출된 토픽 및 그에 따른 단어 조합

Ⅱ. 아이디어 제안	추출된 토픽을 살펴보면 [기능, 디자인, 서비스, 기타] 총 4가지의 토픽들이 추출되었다. 첫 번째 토픽은 기능이다. 기능이라는 토픽 안에 추출된 단어를 살펴보면 온풍, 기능, 청정으로 모두 에어로타워의 기능에 해당되는 단어들이 추출된 것을 알 수 있다. 기능이라는 토픽에 온풍이라는 단어가 가장 높은 확률로 존재하는 것으로 보아, 사람들의 리뷰에 온풍과 관련된 에어로타워의 기능 리뷰가 많이 존재할 것으로 예상된다. 두 번째 토픽은 디자인이다. 디자인이라는 토픽 내에는 디자인, 실내, 베이지라는 단어가 추출되었다. 이 단어들도 모두 디자인에 관련된 단어이다. 세 번째 토픽은 서비스이다. 서비스라는 토픽 내에 배송, 친절, 기사라는 단어들이 추출되었다. 배송과 설치 기사와 관련된 리뷰데이터의 토픽을 잘 추출해 낸 것으로 예상된다. 마지막 토픽은 기타이다. 세 개의 토픽 이외에 다양한 단어들이 추출되었고 이를 기타라고 명명하였다. 추출된 네 가지 토픽의 모델링 결과를 바탕으로 각 토픽에 대한 단어 빈도를 아래 Figure 12., Figure 13.에 시각적으로 표현하였다. 고차원의 토픽을 2차원으로 표현하여 각 토픽별 거리를 쉽게 파악할 수 있으며, 토픽을 선택하였을 때 그 토픽에 존재하는 단어들과, 해당 단어의 사용 빈도수를 시각화 하였다.

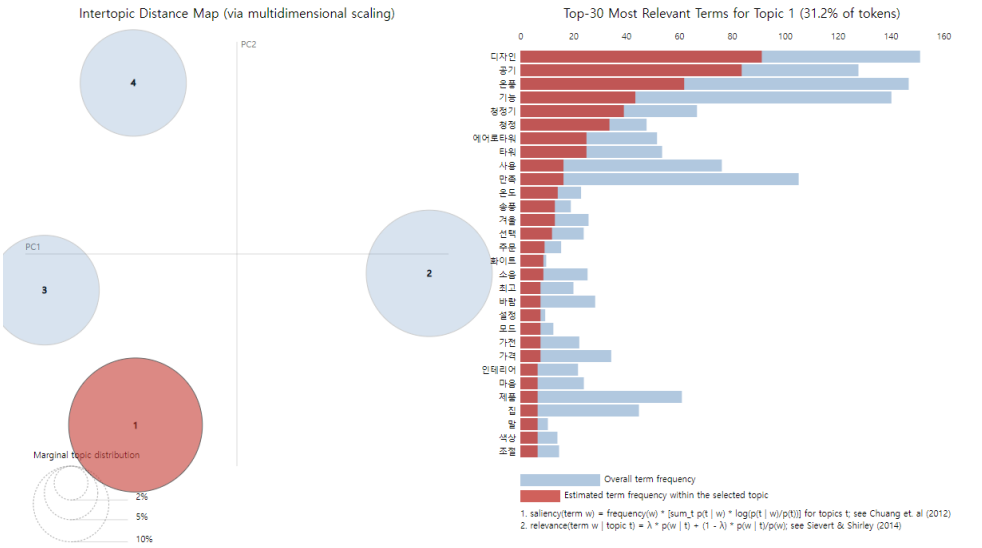


Figure 12. 디자인 토픽에 대한 단어 및 빈도수

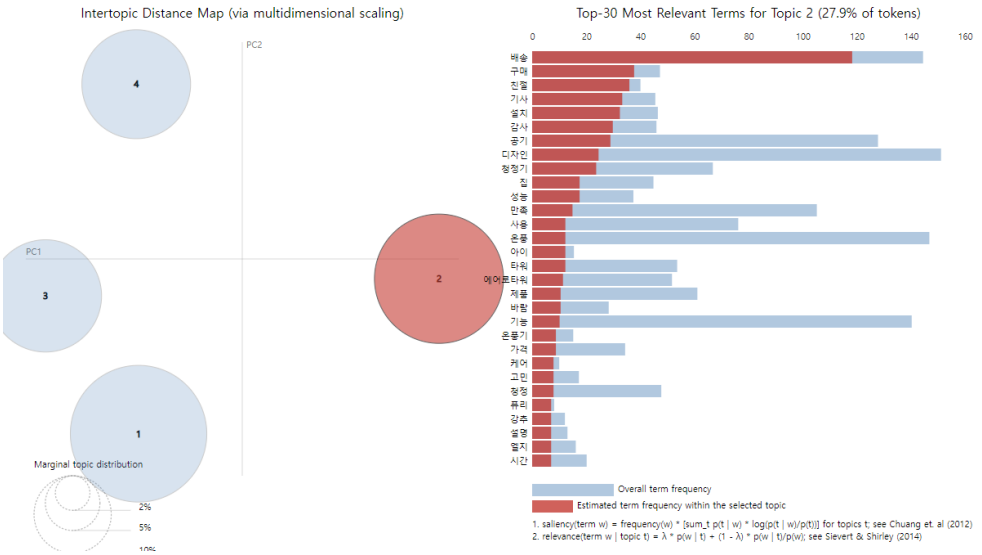


Figure 13. 서비스 토픽에 대한 단어 및 빈도수

II. 아이디어 제안

i. 토픽 별 분석

추출된 최종 4가지 토픽을 살펴보면 소비자들이 LG 전자 에어로타워라는 제품을 평가하고 분석할 때 대부분 기능과 디자인, 그리고 서비스와 관련된 리뷰를 작성하였음을 알 수 있다. Figure 14.는 각 토픽 별 긍정・부정 비율을 보여준다.

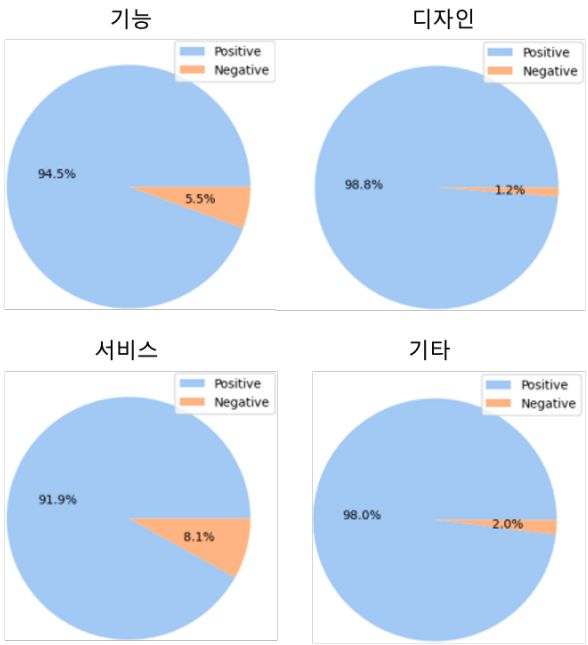


Figure 14. 토픽 별 긍정・부정 비율

각 토픽별 시각화 자료를 보면, 대부분 높은 긍정 비율을 나타내고 있는 것을 볼 수 있다. 다만, 서비스와 관련된 토픽을 보면 부정적인 리뷰의 비중이 가장 높은 것을 볼 수 있다. Figure 15.를 참조하여 이러한 결과는 다른 토픽들에 비해 서비스 토픽 중, '배송' 리뷰가 지배적으로 부정적인 영향을 주는 것을 알 수 있고, 이를 해결한다면 소비자로부터 LG 전자 에어로타워 제품에 대한 긍정적인 평가를 이끌어 낼 수 있을 것이라고 객관적으로 예측할 수 있다.

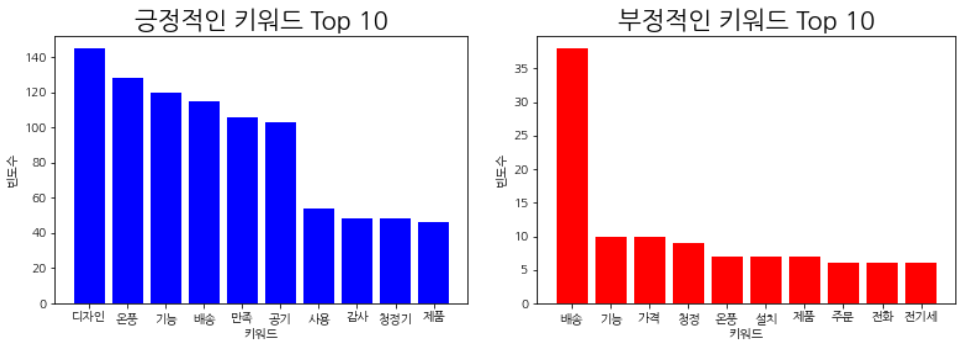


Figure 15. 긍정・부정 상위 10개 키워드

긍정적인 리뷰의 상위 10개 키워드와 부정적인 리뷰의 상위 10개 키워드에서 공통적으로 추출된 키워드들은 [온풍, 기능, 배송, 청정, 제품]이다. 공통으로 추출된 이러한 키워드들은 소비자들이 LG 전자 에어로타워를 사용할 때, 가장 감정을 잘 느끼며, 중요하게 여기는 부분이라고 할 수 있다.

Ⅱ. 아이디어 제안	다음은 긍정 리뷰의 상위 10개 키워드와 부정 리뷰의 상위 10개 키워드에 공통적으로 포함된 키워드들의 편향 비율을 보여준다. * 긍정적인 리뷰 안에 들어 있는 중복을 포함한 단어의 총 개수는 2692개, 부정적인 리뷰 안에 들어있는 중복을 포함한 단어의 총 개수는 315개이다. * 온풍은 긍정적인 리뷰 안에서는 4.75%, 부정적인 리뷰 안에서는 2.22% 차지하므로, 확률적으로 긍정의 감정이 더 강하게 나타난다는 것을 알 수 있다. * 기능은 긍정적인 리뷰 안에서는 4.46%, 부정적인 리뷰 안에서는 3.17% 차지하므로, 확률적으로 긍정의 감정이 더 강하게 나타난다는 것을 알 수 있다. * 배송은 긍정적인 리뷰 안에서는 4.27%, 부정적인 리뷰 안에서는 12.1% 차지하므로, 확률적으로 부정의 감정이 더 강하게 나타난다는 것을 알 수 있다. * 청정은 긍정적인 리뷰 안에서는 1.78%, 부정적인 리뷰 안에서는 2.86% 차지하므로, 확률적으로 부정의 감정이 더 강하게 나타난다는 것을 알 수 있다. 긍정적인 키워드 상위 10개를 살펴보면, 가장 긍정의 감정을 나타내는 키워드는 '디자인'이라고 할 수 있다. 뿐만 아니라, '디자인' 키워드는 긍정적인 키워드 상위 3개에 포함된 '디자인', '온풍', '기능' 중 부정적인 키워드 상위 10개에 포함되지 않은 유일한 키워드로, 가장 긍정의 감정을 나타낸다고 판단할 수 있다. '디자인' 외의 단어들에서는 '온풍', '기능', '배송', '만족', '공기', '감사', '청정' 등의 단어들이 긍정의 감정을 가진 키워드로 나타났다. 긍정의 감정을 가장 잘 나타내는 키워드인 '디자인'을 제품의 홍보에 활용한다면 광고의 효과가 더욱 높아질 것으로 예상된다. 소비자들이 LG 전자 에어로타워 제품에 관심이 있거나 이 제품을 사려고할 때 참고할 시각적 자료에 '공간인테리어 가전'의 키워드를 바탕으로 제품의 디자인을 내세운다면 소비자들의 이목을 더욱 끌 수 있을 것이다. 그러기 위해선, 제품 상세 설명의 '공간인테리어 가전'의 키워드 안에서 LG 전자 에어로타워가 인테리어적인 효과를 더 향상시킬 수 있음을 보여야한다. 그 방법으로는 여러 가정집에 잘 꾸며진 사진을 첨부하는 방법 등이 있을 수 있다. 이런 식으로, '디자인'이란 키워드를 더 강조한다면, 소비자들은 다른 공기청정기들의 디자인과 차별화된 LG 전자 에어로타워의 디자인에 매력을 느낄 것으로 예상한다. 추가적으로, 긍정적인 리뷰의 상위 10개 키워드와 부정적인 리뷰의 상위 10개 키워드에 모두 포함된 '온풍', '기능', '배송', '공기' 키워드 중에서는 긍정의 감정이 부정의 감정보다 더 높은 비율을 차지했던 '온풍'과 '기능' 키워드를 제품의 홍보에 사용하면 좋을 것으로 예상된다. 다음으로, 부정적인 키워드 상위 10개를 살펴보면, 가장 부정의 감정을 나타내는 키워드는 '배송'이다. '배송'뿐만이 아니라 '설치', '주문'의 키워드도 많은 빈도수를 나타내는 것으로 보아, LG 전자 에어로타워의 소비자들이 느끼는 가장 불편함 점은 '서비스'라고 판단할 수 있다. 이 결과에 따라, 배송 서비스를 개선해야 할 것으로 보인다. 추가적으로, '가격', '전기세' 등의 고려할 키워드가 있다. 따라서, 제품을 개선하고자 할 때, 이러한 키워드들을 바탕으로 개선책을 찾아볼 수 있을 것으로 기대한다. 이러한 점들을 이용하여 추후 제안된 아이디어가 상용화 되었을 경우 새로 크롤링되는 사용자 리뷰에 대해 어떤 토픽에 해당하는지 해당 기준을 가지고 분류할 수 있음을 의미한다.

Ⅱ. 아이디어 제안

Ⅱ.2 개발 제품 활용안

제안자는 제안된 아이디어가 활용되는 상황을 지속적으로 크롤링 중인 4개의 온라인 쇼핑 플랫폼 (네이버 쇼핑, 하이마트, 오늘의 집, 다나와) 및 다른 플랫폼 등에서 새로운 리뷰 데이터가 수집 되었을 경우로 가정하였다. 앞의 데이터 수집 항목에서 제안자는 1431개의 문장을 수집하였다. 각각의 문장들에 대한 긍정・부정 척도를 구하였으며, 이를 기반으로 전처리와 모델링을 진행하였다. 전처리된 데이터는 2가지 종류이며, 명사와 동사, 형용사까지 추출된 긍정・부정 예측용 데이터와, 명사만 추출된 토픽 모델링용 데이터로 구성된다. 이 중 전자를 사용하여 구한 문장들의 긍정・부정 비율은 Figure 16.과 같다.

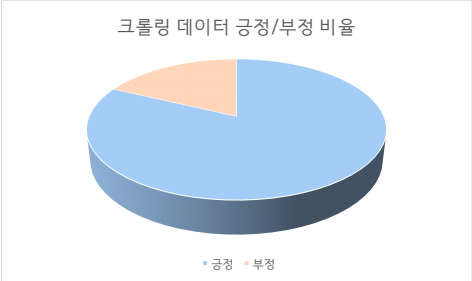


Figure 16. 1431개의 문장에 대한 크롤링 데이터 척도 비율 (~= 7:3)

a. 감성 분석 모델의 필요성

새로 들어온 문장에 대해 긍정・부정을 분류해주어야 제안된 아이디어 제품에 활용될 수 있다. 이를 위해 전처리된 데이터를 기반으로 한 감성 분석 모델을 만들었으며, 제안자는 XGBoost 모델을 이용한 트리 구조 기반 앙상블 모델을 사용하였다.

b. 트리 기반 모델

먼저 트리 기반 모델에 대해 설명한다. 트리 기반 모델이란, 이진 분류를 시행하는 여러 개의 노드를 만들고 그 노드들을 계층구조로 쌓아 올린 후, 입력된 특징에 대한 노드들의 이진 분류 결과를 종합하여 결론을 내는 모델이다.

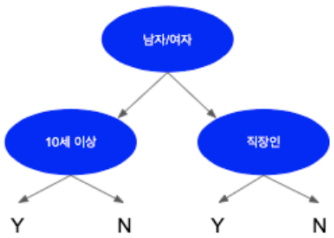


Figure 17. 2개 계층의 트리 구조

이때, 트리 기반 모델은 Entropy 값을 최소화하는 것을 목표로 이진 분류를 행한다. 데이터 분석에서 Entropy란, 얼마나 정보에 대해 알고 있는지, 혹은 모르는 데이터가 얼마나 있는지에 대한 척도이다. 정보를 많이 알수록, 모르는 데이터가 적을수록, 그리고 새로운 데이터에 의한 영향이 적을수록 Entropy가 낮아진다. 트리 기반 모델이 Entropy를 최소화하려고 한다는 것은, 최대한 정보를 많이 알게 된 상태에서 결정을 하게 된다는 것을 의미한다. 이는 정확도의 향상으로 이어진다.

Ⅱ. 아이디어 제안	c. 앙상블 모델: XGBoost 데이터 분석에서 앙상블이란, 투표를 통해 다수결의 의견을 채택하듯이 여러 모델의 결과를 반영하여 종합적으로 의사결정을 하는 것이다. 예를 들어 2가지 모델로 값을 예측할 때, 그 값이 정답이라고 볼 수 있는지에 대한 확률을 구하고 비교할 수 있다. 이러한 확률을 기반으로 의사결정을 한다면, 이는 2가지 모델의 앙상블이라고 할 수 있는 것이다. 제안된 아이디어에서 사용된 앙상블 모델은 XGBoost 모델이다. XGBoost 모델은 Gradient Boosting 기법이 병렬 시행되도록 개발된 트리 기반 앙상블 모델이다. 이러한 모델은 잘못 예측한 값들에 대해 가중치를 부여하여 다시 학습을 시킨다는 특징을 가진다. 이것을 다수의 모델에 적용하고, 각각 다른 성능을 가지고 있는 모델들을 고려하여 가중치를 부여한다. 이것을 반복하여 점진적으로 학습시킨다. XGBoost 기법은 병렬 시행이 지원되므로 다른 앙상블 모델들에 비해 연산 속도가 빠르다는 장점이 있다. 또한, 데이터의 Outlier에 의한 영향이 적거나 Scaling이 필요하지 않는 등, 트리 기반 모델의 여러 장점과 함께 빠른 연산 속도를 갖고 있어 현대 데이터 분석에서 종종 채택되는 모델 중 하나이다. d. 리뷰의 긍정・부정 판단 감성 분석 모델의 개발 1431개의 전처리된 문장 데이터를 10개의 Cross-Validation 세트로 나누었고, 정확도를 기준으로 모델 구축 및 평가를 진행하였다. Logistic Regression, Random Forest, XGBoost, Sequential 딥러닝 모델 등 여러 모델을 학습하였고, 그중 결과가 가장 우수한 XGBoost 모델의 정확도 결과는 Table 3.과 같다. Cross Validation 1: 86.81% Cross Validation 2: 86.01% Cross Validation 3: 86.71% Cross Validation 4: 87.41% Cross Validation 5: 83.22% Cross Validation 6: 82.52% Cross Validation 7: 79.02% Cross Validation 8: 83.92% Cross Validation 9: 83.22% Cross Validation 10: 79.72% Table 3. XGBoost 모델의 Cross Validation 정확도 위의 결과를 종합하면, 학습한 예측 모델은 평균적으로 83.61%의 정확도를 갖는다는 것을 알 수 있다. 이를 리뷰의 긍정・부정 판단 모델로 사용하여 새로 들어오는 리뷰 데이터에 대해 긍정・부정을 판단하였다. 이 정확도는 제품 판매 기간이 길어질수록 쌓이는 리뷰 데이터들의 수에 비례하여 증가한다.
Ⅲ. 현장 활용 방안	Ⅲ.1 상용화 방안 GUI (Graphic User Interface) 개발은 자동화에 초점을 두고 진행되었다. 프로그램 진행 과정은 다음과 같다. 분석가가 입력한 검색 키워드에 대해 분석가가 선택한 온라인 쇼핑 플랫폼에서 사용자 리뷰를 크롤링한다. 이렇게 수집된 사용자 리뷰를 토큰화된 데이터로 전처리한다. 여기서 두 가지로 프로그램이 갈리는데, 첫 번째로, 학습된 감성 분석 모델을 이용해 사용자 리뷰의 긍정・부정을 판단한다. 이후 긍정・부정 별 상위 키워드를 GUI 상에 시각화한다. 두 번째로, LDA를 이용해 토픽을 추출한 후 토픽별 긍정・부정 비율을 GUI 상에 시각화한다. 프로그램 진행 과정을 도식화 하여 Figure.18에 나타내었다.

Ⅲ. 현장 활용 방안

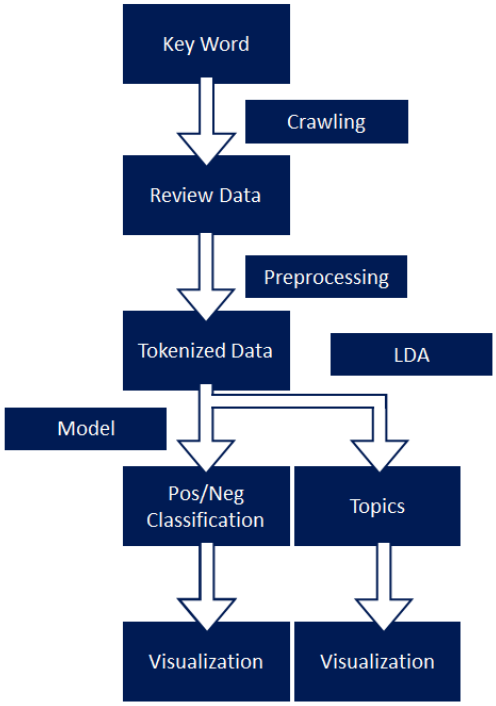


Figure 18. 프로그램 흐름 도식화

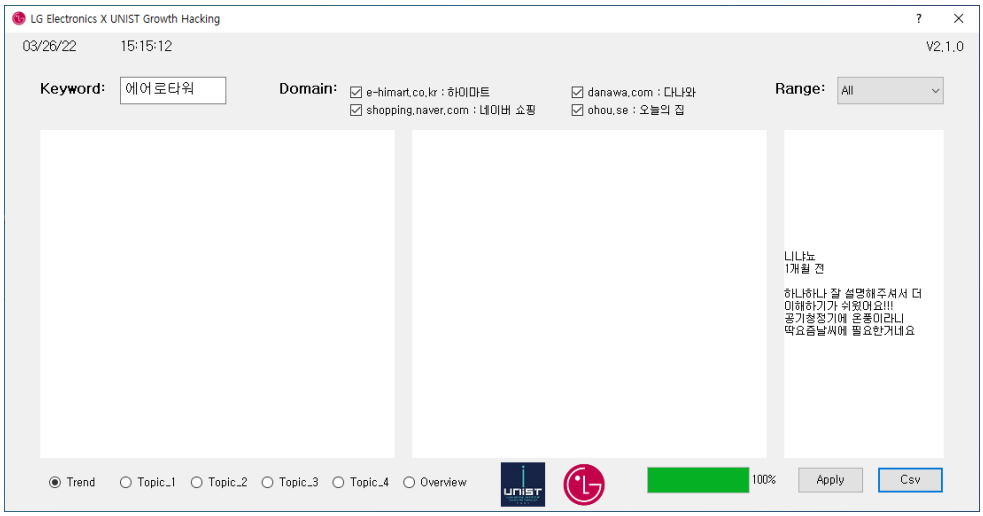


Figure 19. GUI (v2.1.0) 초기 화면

GUI 상에 분석가가 입력할 수 있는 것은 [1. 검색 키워드, 2. 크롤링 대상 온라인 쇼핑 플랫폼 도메인, 3. 분석할 데이터 시간 범위]이다. 키워드는 크롤링할 데이터의 검색어이며, 도메인은 4가지 예시로 구성되어 있으며, 추가 가능하다. 또한 시간 범위는 [1주, 2주, 1달, 6달, 전체]로 구성되어 있다. 키워드로 '에어로타워'를 입력한다면 선택된 크롤링 대상 온라인 쇼핑 플랫폼에서 '에어로타워'를 검색하고 지정 시간 범위내의 데이터를 자동으로 수집한다. 이후, 해당 데이터에 대해 프로그램 진행 과정에서 언급한 모든 작업들이 자동적으로 수행된다. 즉, 데이터 크롤링, 전처리, LDA 기반 토픽 모델링, XGBoost 모델을 통한 긍정·부정 예측, 긍정·부정 데이터에서 자주 등장한 단어 Top 10, 토픽별 긍정·부정 비율을 자동적으로 계산한다. 또한, 계산한 데이터를 직관적으로 이해할 수 있도록 시각화하여 GUI 상에 띄워 주는 기능을 개발하였다.

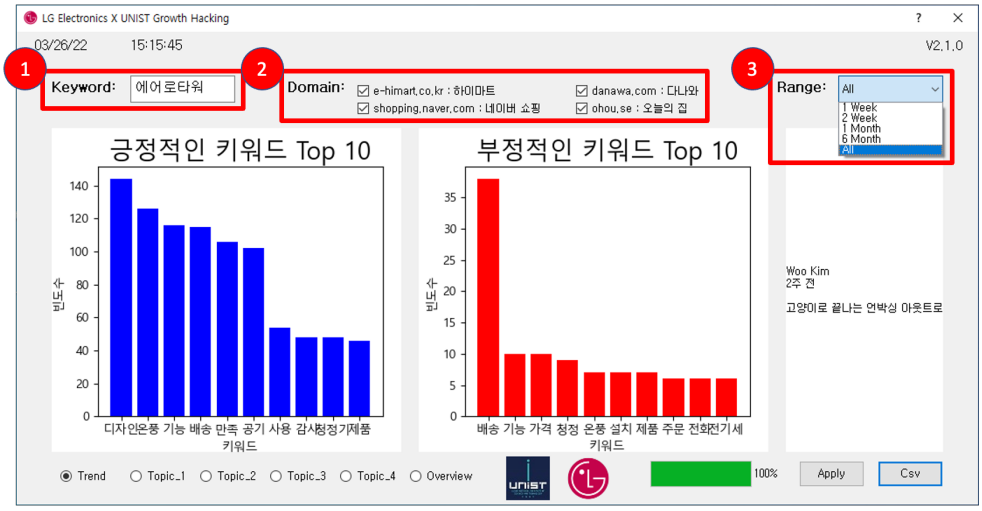


Figure 20. GUI (v2.1.0) 사용자 입력 항목 3가지

GUI 상에 나타나는 시각화 자료로는 [4. Left Display, 5. Right Display, 6. 실시간 유튜브 댓글]이 있다. 두 개의 Display에 어떤 시각화 자료를 보여줄지는 7. Display Switch 기능에 종속되는데, 이 기능은 [Trend, Topic_1, Topic_2, Topic_3, Topic_4, Overview]의 옵션으로 구성되어 있다. Figure 21.은 Trend 옵션이 활성화 되어있을 경우 나타나는 시각화 자료로서, [긍정적인 키워드 Top 10, 부정적인 키워드 Top 10] 시각화 자료가 보여진다. Topic_N 옵션은 [각 토픽에 대한 긍정・부정 비율, 각 토픽에 따른 키워드 분석]을 시각화한다.

Ⅲ. 현장 활용 방안

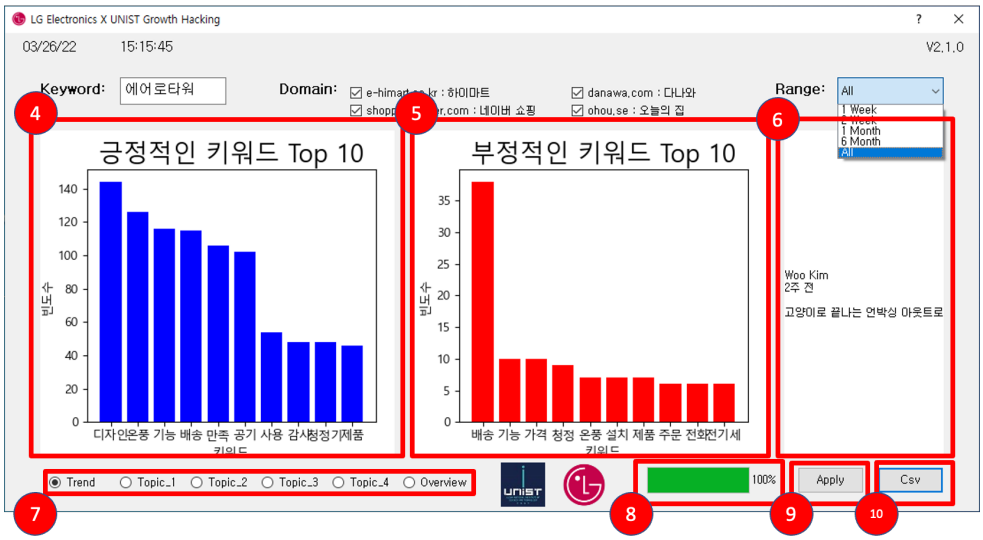


Figure 21. GUI (v2.1.0) 시각화 자료 및 기능 버튼

Figure 22.는 Overview 옵션이 활성화 되어있을 경우 나타는 시각화 자료로서, [모든 토픽의 긍정・부정 비율 요약, 토픽별 세부 키워드] 시각화 자료가 보여진다. 6. 실시간 유튜브 댓글 기능은 입력된 키워드에 대한 유튜브 영상 댓글을 수집하며, 여러 게시물을 수작업으로 찾는 비용을 줄이고 실시간 반응을 빠르게 파악하기 위해 고안된 기능이다. 입력한 검색 키워드를 유튜브에 검색했을 경우 나오는 모든 자료에 대한 댓글을 크롤링으로 수집하여 작성자와 작성 시간, 댓글을 텍스트로 요약하여 보여준다. 이 기능은 실시간으로 자료를 수집한 후 시각화하며, 새로 고침 주기는 1초이다.

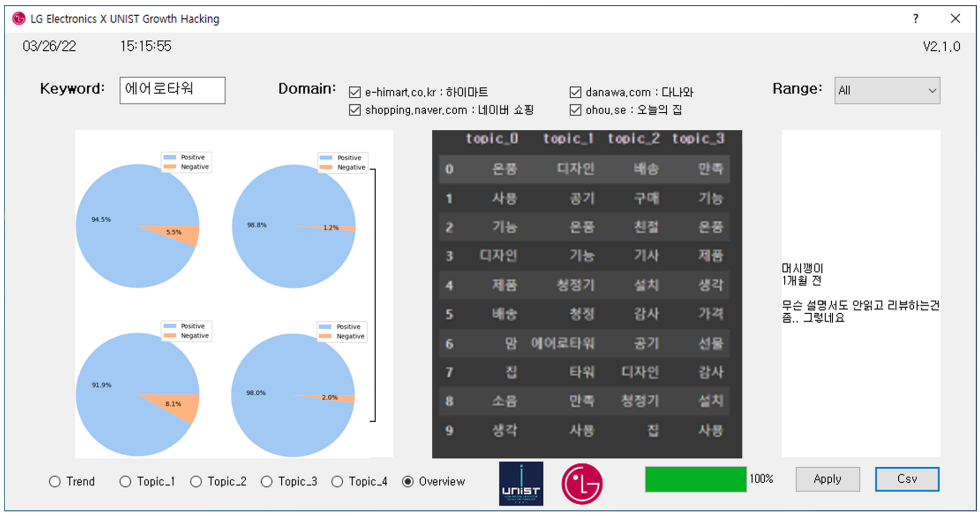


Figure 22. GUI (v2.1.0) Overview 옵션에 따른 시각화 자료

하단에 위치한 8. Progress Bar는 분석가가 요청한 작업에 대한 진행도를 시각화하여 보여주며, 9. Apply 버튼은 분석가가 선택한 옵션에 따른 시각화 자료의 새로 고침을 요청한다. 참고로 6. 실시간 유튜브 댓글 기능은 이 작업만을 별도로 진행하는 쓰레드를 배정하였으므로, Apply 클릭 유무에 상관없이 멀티태스킹으로 작동한다. 10. CSV 버튼은 프로그램 작동 시점부터의 모든 수집 및 분석한 데이터를 쉼표로 구분된 값을 가지는 CSV 포맷으로 저장하며, 추후 추가적인 분석 및 다른 분석 프로그램과의 연동을 위한 기능이다.

Ⅲ. 현장 활용 방안