The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 19, No. 6, pp.1-8, Dec. 31, 2019. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2019.19.6.1
JIIBC 2019-6-1

빅데이터 분석을 활용한 가짜 리뷰 필터링 시스템 ADDAVICHI

Development of Filtering System ADDAVICHI for Fake Reviews using Big Data Analysis

정다비치*, 노영주**

Davichi Jeong*, Young-J. Rho**

요 약 최근 '바이럴 마케팅' 으로 인해서 홍보에만 치중하는 블로그 게시물 등으로 인해 소비자의 불신이 깊어졌다. 또한, 이용후기를 거짓으로 작성하거나, 과장 확대하는 등의 마케팅 사업은 신문이나 TV 광고에 비해 가격이 저렴하면서도 효과가 커 각광받는 사업 중 하나로서 광고비 규모는 2016년 기준 '3조 3941억'으로 주요 광고수단으로 자리잡고 있다. 이러한 '바이럴 마케팅'으로부터 정보를 걸러주는 도구가 필요한 인터넷 환경이 되었다. 본 논문에서 제시하는 가짜 리뷰 필터링 어플리케이션 ADDAVICHI는 사용자가 '이벤트', '맛집' 등의 컨텐츠를 검색하면 블로그 키워등, 총 검색수, 신뢰도, 만족도 등을 추출하고 분석하여 제시한다. 신뢰도는 블로그에 있는 광고게시물 수와, 전체 게시물 수를 보여주고, 만족도는 신뢰도에서 걸러진 청정 게시물을 긍정 게시물과 부정게시물로 나눠서 보여준다. 마지막으로 키워드는 긍정 게시물에서 나온 리뷰 상위 세 단어 리스트를 보여준다. 이러한 방법으로 사용자가 광고 글로부터 벗어나서 정보를 해석할 수 있도록 지원한다.

Abstract Recently, consumer distrust has deepened due to blog posts focusing only on public relations due to 'viral marketing'. In addition, marketing projects such as false writing or exaggerated use of the latter phase are one of the most popular programs in 2016 as they are cheaper and more effective than newspaper and TV ads, and the size of advertising costs is set to be a major means of advertising at '3 trillion 394.1 billion won. From this 'viral marketing,' it has become an Internet environment that needs tools to filter information. The fake review filtering application ADDAVICHI presented in this paper extracts, analyzes, and presents blog keywords, total number of searches, reliability and satisfaction when users search for content such as "event" and "taste restaurant." Reliability shows the number of ad posts on a blog, the total number of posts, and satisfaction shows a clean post with confidence divided into positive and negative posts. Finally, the keyword shows a list of the top three words in the review from a positive post. In this way, it helps users interpret information away from advertising.

Key Words : Mobile, text mining, significant advertising review, content positive/negative

*회원, 한국산업기술대학교 컴퓨터공학부
**정회원, 한국산업기술대학교 컴퓨터공학부(교신저자)
접수일자: 2019년 10월 24일, 수정완료: 2019년 11월 24일
게재확정일자: 2019년 12월 6일

Received: 24 October, 2019 / Revised: 24 November, 2019 / Accepted: 6 December, 2019
**Corresponding Author: yrho@kpu.ac.kr
School of Computer Engineering, Korea Polytechnic University, Korea

Ⅰ. 서 론

바이럴 마케팅(viral marketing)이란 위키피디아(wikipedia)에서는 소셜미디어, 이메일 등 매체를 통해 대중이 자발적으로 메시지를 퍼트리는 마케팅 방법으로 설명하고 있다[1]. 현재 검색어 표출, 소개글 작성 등의 바이럴 마케팅 업체가 성행함에 따라 검색어나 블로그, 카페는 물론이고 유명 소셜 네트워크 (SNS) 등에서도 순수하게 관련 정보를 제공하는 듯한 페이지를 운영하며 홍보효과를 극대화 하려는 방식을 활용하고 있다. 하지만 바이럴 마케팅의 경우 홍보에만 치중하다 보니 이용후기를 거짓으로 작성하거나 과장, 확대하는 경우도 많은 것으로 알려졌다. 신문이나 TV 광고에 비해 가격이 저렴하면서도 효과는 더 크다보니 홍보수단으로 각광받고 있다.

뉴스워치 보도[2]에 따르면 예전과 달리 인터넷의 매체별 광고비 규모는 '3조 3941억'으로 매우 높고, 매체별 개별로 가게 되면 온라인 광고 수익이 가장 높게 나타나고 있다. 또한 ㈜ 애드픽네트워크 관계자는 2017년에 비해 2018년 바이럴 마케팅 광고비가 더욱 더 많아질 것이라 예측했다. 현재는 어린 10대 층만이 이러한 것에 영향을 미치는 것이 아닌, 20대 ~ 50대 연령까지 바이럴 마케팅을 통해 소비가 이루어지는 경향이 있다.

이러한 경향으로 저렴하고 효과가 큰 주요 마케팅방법으로 바이럴 마케팅이 활용되고 있으며, 이와 함께 이용후기를 거짓 작성하거나 과장 확대하는 등의 방법이 늘어나 이용소비자의 불신도 함께 깊어지고 있다.

본 논문에서는 소비자 불신을 개선하기 위한 방법으로 ADDAVICH 시스템 연구개발을 진행하였다. 텍스트 마이닝을 위한 기법들은 다양한 방법들이 있으나[3], ADDAVICH에서는 웹 크롤링을 이용하여 수집한 블로그 게시물을 기반으로 BRNN[5]과 KoNLPy[6]를 이용하여 블로그 총 검색 수, 신뢰도, 만족도, 키워드를 추출하고 분석하여 제시하는 방법을 적용하였다.

Ⅱ. 관련 기술 조사 및 적용

1. 웹 크롤링

웹 크롤링은 Web 인덱싱을 체계적으로 하는 browser라고 할 수 있으며, 각 웹 검색기에서 대표적으로 사용된다[19]. 네이버나 다음에서 블로그 리뷰 페이지를 가져오기 위한 웹 크롤링에서는 Python 라이브러리 BeautifulSoup[7]가 사용된다.

BeautifulSoup으로 크롤링 한 리뷰는 Word2Vec[8]을 통해 임베딩 벡터로 변환되고, 다시 안드로이드 시스템에서 사용할 수 있는 PHP 화면으로 변환하는 방법이 사용된다 (그림 1 참조).

```
import requests
import re
from bs4 import BeautifulSoup
from collections import OrderedDict
from itertools import count
import db_Insert

def naver_content_cralwler(url):
    hrd = {'User-Agent':'Mozilla/5.0', 'referer':'http://naver.com'}
    param = { 'where': 'post' }

    try:
        response = requests.get(url, params=param, headers=hrd)
        soup_temp = BeautifulSoup(response.text, 'html.parser')
        area_temp = soup_temp.find(id='screenFrame')
        url2 = area_temp.get("src")
    except:
        try:
            area_temp = soup_temp.find(id='mainFrame')
            url3 = area_temp.get("src")
            url4 = "https://blog.naver.com" + url3
            return url4
        except:
            return ""
```

그림 1. 파이썬을 이용한 네이버 웹 크롤링 코드 예
Fig. 1. Naver Web crawling code using Python

2. BRNN을 이용한 텍스트 감성분석

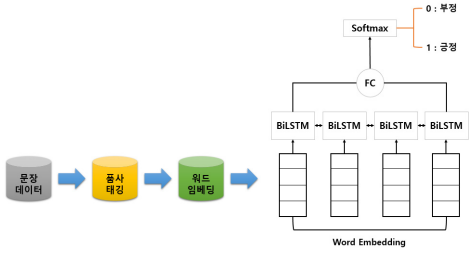


그림 2. BiLSTM 모델 구조도
Fig. 2. Schematic Model of BiLSTM[

텍스트 감성분석은 그림 2와 같이 BRNN인 BiLSTM(bidirectional long short-term memory) 모델을 이용하는 방법이 Doc2Vec Logistic Regression이나 Term-existance Naive Bayes 보다 상대적으로 높은 정확도를 보인 실험적 보고가 있다[9].

BiLSTM 감성분석은 정답이 있는 네이버 리뷰 데이터 15만건에 대해서 품사 태깅 하고, 태깅한 단어들에 대해 Word2Vec을 이용해 학습시켜 임베딩 벡터로 변환한다. 그리고 단어 벡터들을 BilSTM에 넣어서 양쪽 끝 state들

에 대해서 fully connected layer와 Softmax 함수[10]를 이용해 분류한다.

이 방법에는 KoNLPy[5] 패키지가 필요하며, 추가로 tensorflow[11]와 gensim[12] 패키지도 필요하다. 이러한 패키지들을 이용해서 네이버 블로그 글에 대한 긍정/부정 분석을 하고, 이를 통해 청정 게시글에 대한 신뢰도 분석이 가능하다.

3. KoNLPy를 이용한 키워드 분석[14][21]

KoNLPy는 한국어 자연어 처리를 위한 Python 패키지이다[5]. 키워드 추출을 위해서는 KoNLPy, collections 파이썬 모듈[13]을 사용한다.

KoNLPy의 Twitter 객체에서 nouns 함수를 통해서 text에서 명사만 분리/추출하며, 객체 안의 명사 중 빈도수 별로 정렬하여, 큰 명사부터 순서대로 입력받은 정수 개수만큼 저장되어있는 객체를 반환한다. 이를 통해 청정 게시글에 대한 상위 키워드를 추출한다.

4. 블로그 정보를 이용한 광고 필터링

광고 필터링은 네이버 블로그 상의 정보를 기준으로 이루어진다. 네이버 블로그 상의 정보 중에서 필요한 정보는 컨텐츠 관련 다른 게시물 수, 컨텐츠 관련 사진 수, 후원, 무료, 광고, 업체, 제공 등의 어휘 사용, 컨텐츠 관련 글 내용 등의 내용을 종합시켜 해당 블로그의 정보를 판단한다. 이를 통해 광고 게시글 또는 청정 게시글에 대한 정보를 얻는다.

5. Selenium 파이썬 라이브러리를 이용한 크롤링

네이버나 다음에서 제공하는 리뷰 페이지는 다음 페이지를 넘기기 위해 웹 브라우저 상의 액션이 이루어져야 모든 리뷰 목록을 볼 수 있다. 이러한 액션을 자동으로 수행하는 방법으로 Selenium[14]이 있다. Selenium은 Python 환경에서 수행되며 상품 URL을 통해 웹 브라우저를 실행시키고 상품 리뷰를 수집한다.

이렇게 크롤링 한 리뷰는 한나눔 형태소 분석기 파이프라인 구조를 이용하여 형태소 분석을 수행한다.

6. 한나눔 형태소 분석기를 이용한 형태소 분석

한글의 형태소 분석은 음소로부터 문장까지 그림 3과 같은 구조로 이루어진다. 형태소 분석기에는 한나눔 형태소 분석기[15] 이외에도 꼬꼬마[16], MeCab-Ko[17] 등이 있으나 한나눔 형태소 분석기를 적용하였다.

한나눔 분석기는 그림 4와 같이 차트 기반의 형태소 분석 기법을 사용하고 있으며, v0.8.4 버전 기준으로 데이터는 총 언어자원은행에 등록된 기준으로 69개의 문서에 포함된 총 56,339개의 문장, 768,708개 어절이 있으며 품사 정확도는 89% 이다.

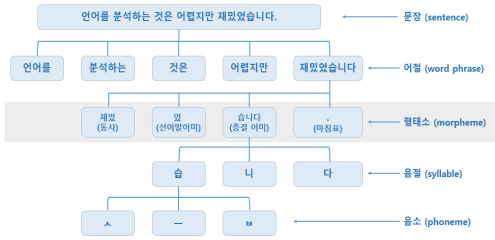


그림 3. 한글 형태소 분석 구조도
Fig. 3. Structure of Hangul Morphological Analysis

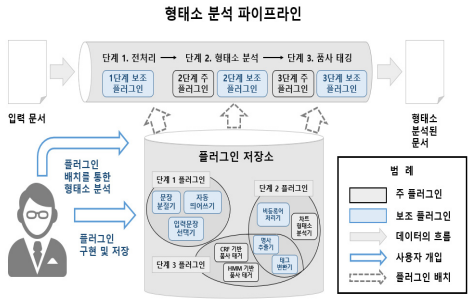


그림 4. 데이터 형태소 분석 파이프라인
Fig. 4. Data Morphological Analysis Pipeline

7. 패턴 추출 과정

구조번호	품사패턴	구조번호	품사패턴
1	NV	4-1	VN
2-1	NZV	4-2	N1N2
2-2	NZN	5-1	ZVN
3	N1N2N3	5-2	ZN1N2

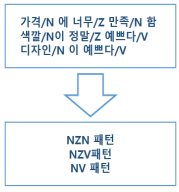


그림 5. HMM 기반의 품사 태깅 - 패턴 추출 예
Fig. 5. Example of Generic tagging based on HMM

한나눔 형태소 분석기는 HMM(Hidden Markov Model) 기반의 품사 태깅 방식을 채용하고 있으며, 그림 5는 8가지 품사 패턴 구조와 패턴의 추출 예시를 담고 있다. 패턴 추출 과정은 상품의 특징 단어를 기준으로

±2 위치의 단어를 가져오는데, 여기서 상품의 특징 단어는 상품의 가격, 디자인 등이 될 수 있다. 가져온 단어를 8가지 패턴과 비교하여 8가지 패턴에 해당한다면 감정 분석 과정으로 넘어간다.

8. 긍정 또는 부정 어휘 추출

추출된 패턴에 포함하는 감정어가 긍정, 부정, 중립 중에서 어떤 의미인 지를 판별하기 위해 감정어 사전을 사용한다. 감정어 사전의 구축은 영어권에서 대표적 감성사전 중 하나인 SentiWordnNet (SWN)[18]을 사용한다..

감정어 사전은 긍정적 의미인 경우 1점, 부정적 의미인 경우 -1점을 두었으며, 중립적인 경우 0점을 부여한다. 그림 6은 긍정 어휘 추출 예를 보여준다.

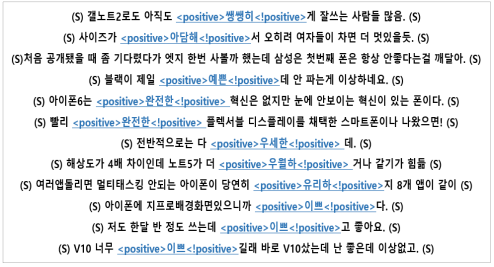


그림 6. IT 리뷰 코퍼스의 긍정 어휘 추출 예
Fig. 6. Examples of positive vocabulary extraction of IT review copersons

9. SWN 어휘 한국어 사전 선별

SWN에 수록된 어휘는 총 117,659 개이다. 이 중에서 긍정 및 부정 수치를 각각 0.5를 기준으로 하여 그 이상인 단어들만 추출한다면, 감성어휘는 대략 13,000 개 정도 나오게 되는데 이 어휘를 구글 번역기를 이용해 번역하고, 중복 제거 해 대략 9,000 개 정도의 감성어휘 목록을 선별할 수 있게 된다. 그림 7은 구글 번역기의 문제점을 해결하기 위해 마련한 극성, 품사 정보 재정의 자료이다.

10. 극성, 품사 재정의, 규모

그림 7에서 첫 번째 그림은 강한긍정 / 긍정 / 중립 / 부정 / 강한부정 / 상대극성 / 강한 상대극성의 7가지로 나눈다. 두 번째 그림은 명사 / 형용사 / 동사 / 부사 / 여러 단어 및 구 의 5가지 품사 정보로 나눈다.

이렇게 감성어휘를 재정의해서 구분하게 되면, 표 1에서와 같이 총 3,665개의 감성사전 규모가 된다.

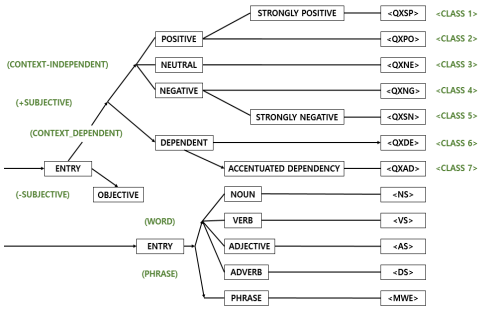


그림 7. 한국어 감성어휘 극성, 품사 정보
Fig. 7. Extreme property, Verbal Information of Korean Emotional Vocabularies

표 1. SWN을 기반으로 구축된 감성사전 규모
Table 1. Scale of Emotional Dictionary based on SWN

TAG	NOUN	ADJECTIVE	VERB	ADVERB	TOTAL
QXSP	31	13	6	1	51
QXPO	650	275	104	52	1,081
QXNE	22	8	2	0	32
QXNG	1,425	339	221	47	2,032
QXSN	50	11	5	0	66
QXDE	222	79	70	19	390
QXAD	9	2	1	1	13
Total	2,409	727	409	120	3,665

11. 블로그 관련 부가 정보 활용

네이버 블로그, 다음 티스토리 상에는 포스트 된 수, 팔로워 된 수 등의 정보가 제공된다. 이러한 정보와, 다음과 같은 여러 정보를 종합해 일차적으로 가중치를 부여한다.

- 검색 조회수 (포스트, 팔로워)
- 컨텐츠 관련 다른 게시물 수
- 통합검색에서 블로그 영역

12. 광고성 블로그 필터링 방법

광고성 블로그 글은 댓글과 블로거 사이의 관계 등을 파악해 신뢰성과 명성을 수치화 할 수 있다. 여기서 광고성 블로그는 이 수치가 떨어지면서 자연스럽게 가중치가 낮아지게 되는데, 예를 들면, 옛날에 쓴 글을 지운 뒤 다시 쓰는 것을 반복한다던가, 댓글부대를 동원하는 것도 광고성 블로그를 판별하는 기준 중 하나이다. 실제 다이닝 코드[20]라는 빅데이터 기반 맛집 판별 어플에서는 이 기법을 사용하여 광고성 블로그를 필터링 한다.

추가적인 필터링 방법으로는 “이 블로그 글은 무료로 제품을 제공 받고 쓰는 제품입니다” 라는 글 형식의 블로그나, 추천인 아이디를 마지막에 써서 사용자를 유도하는 형식의 글은 필터링을 한다. 또한 지나친 고화질의 사진,

같은 단어의 반복 등의 패턴에서도 광고성 리뷰를 찾는다.

III. 본 론

1. ADDAVICHI (가짜 리뷰 필터링 APP) 기능

ADDAVICH 시스템은 그림 8과 같은 시작 화면과 아이콘으로 시작한다.

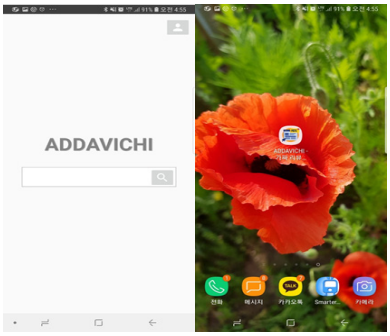


그림 8. 어플리케이션 시작 화면과 아이콘 화면
Fig. 8. Application start screen and icon screen

그림 9. 어플리케이션 로그인, 회원가입 화면
Fig. 9. Application login, membership registration screen

본 논문에서 소개하는 어플리케이션은 사용자가 '플러그인'이나 '다른 어플리케이션'을 통해 들어가는 것이 아니라 직접 어플리케이션에 들어가서 사용해야 한다. 향후 WEB 버전도 개발 예정이나, 본 논문에서는 어플리케이션 버전만 소개한다.[22]

그림 9에서 보이는 '검색 창'을 이용하여 원하는 컨텐츠를 검색할 수 있으며, 홈 버튼을 누르면 초기화 상태로 돌아간다. 그림 10, 11은 검색 화면 예시를 보여준다.

회원가입 중에는 중복 제한 기능이 있다. 또한 아이디 체크가 완료되면 아이디를 수정할 수 없다. 화면은 로그인 화면, 회원 가입 화면으로 나누어진다.

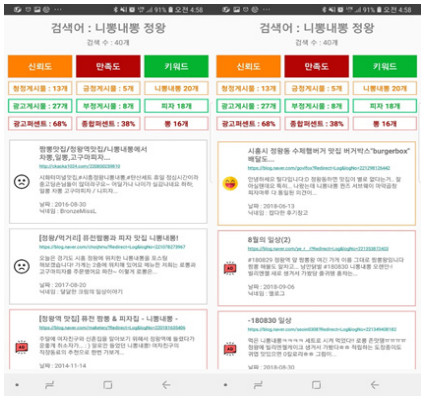


그림 10. 검색 화면 1
Fig. 10. Search screen 1

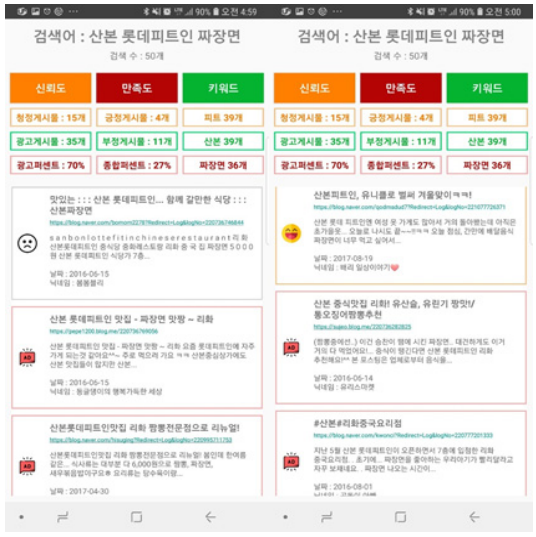


그림 11. 검색 화면 예2
Fig. 11. Second Example of Search Screen

회원가입 화면을 누를 시 회원가입 창으로 나누어지며, 로그인 이후에는 어플을 종료했다 다시 로그인 해도 자동 로그인 처리가 된다.

본 논문의 어플리케이션에서 처음 검색을 하면, 긍정 / 부정 / 광고 이미지가 표시되는 대신, 로딩 이미지가 표시 된다. 또한 신뢰도 / 만족도 / 키워드 역시 빈칸으로 나오게 된다. 처리가 완료되면 표출 되는 방식이다.

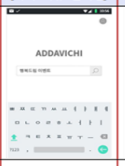
현재는 네이버 블로그 게시글로만 검색을 진행했다.

화면을 설명하자면, 처음 부분에는 검색어, 검색 수, 중간 부분에는 신뢰도, 만족도 키워드, 끝 부분에는 '네이버 블로그' 게시글 내용이 표시된다.

2. 다른 컨텐츠와의 비교

표 2는 다른 시스템과 비교한 도표이다.

표 2. 다른 컨텐츠와의 비교 차트
Table 2. Comparison Chart with Other Contents

	ADDAVICHI [가짜 리뷰 필터링]	다이닝 코드	소셜 매트릭스	PulseK
개요				
구분	본 프로젝트 개발예정	국내 컨텐츠	국내 컨텐츠	국내 컨텐츠
핵심가치	사이트 전체에 대한 신뢰도/만족도 제공	맛집 랭킹 제공	이슈 검색어 제공 긍정/부정 제공	이슈 검색어 제공 긍정/부정 제공
대표 제공 서비스	만족도 제공 광고량 제공 커스텀 비교서비스	맛집 랭킹 제공 광고 필터링 주변 맛집 확인	연관 키워드 순위 감성 키워드 순위 주간 급증 키워드	이슈어-연관어 긍, 부정 언급량 TOP 10 이슈어
지원 플랫폼	모바일 [Mobile] 웹 [향후 지원 예정]	모바일 [Mobile] 웹 [Web]	웹 [Web]	웹 [Web]

다이닝 코드는 '맛집' 이라는 한정된 컨텐츠를 가지고 있고, '맛집' 이라는 컨텐츠 특성 상 광고량에 대한 정보가 필요 없다. 또한 소셜매트릭스, PulseK는 연관 키워드, 긍정/부정에 대한 정보는 알려주지만 '광고 데이터' 에 대한 필터링이 없어 가짜 리뷰를 많이 포함했다면 정확한 긍정, 부정 여부를 확인할 수 없다.

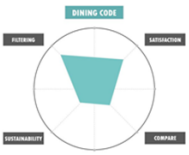
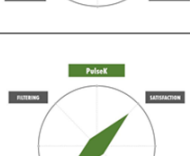
반면에 본 논문에서 소개하는 ADDAVICHI (가짜 리뷰 필터링)은 연관 키워드, 긍정/부정에 대한 정보를 알려줄 뿐만 아니라 '광고 데이터' 에 대한 필터링, '광고량' 에 대한 정보를 화면에 명시적으로 표시하고, 랭킹을 사용자가 적합화하여 데이터로 나타낼 수 있는 차별점이 있다. 표 3은 이런 관점에서 다른 컨텐츠와 비교 분석한 도표이다.

다이닝 코드의 경우 필터링, 만족도 부문에서는 높은 지표를 보이나, 시기별 '맛집' 을 보여주는 기능인 지속성, 한꺼번에 비교해주는 '비교' 기능은 없다.

소셜 매트릭스, PulseK의 경우 키워드에 따라 긍정/부정 추이를 알 수 있는 만족도, 시기별 '컨텐츠'의 긍정/부정 추이를 알 수 있는 지속성에서는 높은 지표를 보이나, 필터링, 비교 부문 기능이 없어 낮은 지표를 보인다.

본 논문에서 소개하는 ADDAVICHI (가짜 리뷰 필터링)은 모든 부문에서 고른 지표를 보여 이 부문이 다른 프로그램과의 차별점이다.

표 3. 다른 컨텐츠와의 비교 차트
Table 3. Comparison Chart with Other Content

	[ADDAVICHI] 1. FILTERING [필터링] - 가짜 리뷰 글에 대한 필터링 후, 광고 선별합니다. 2. SATISFACTION [만족도] - '컨텐츠' 에 대한 긍정/부정 표출, 만족도 계산합니다. 3. SUSTAINABILITY [지속성] - 시기별로 '컨텐츠' 에 대해 차트로 표출됩니다. 4. COMPARE [비교] - 사용자 커스텀마이징 '비교' 로 점수 환산 가능합니다. 5. 정리 – 4가지 지표 모두 고른 점수 확득했습니다.
	[다이닝 코드] 1. FILTERING [필터링] - 자체 기술로 '맛집 가짜 리뷰' 글 필터링 됩니다. 2. SATISFACTION [만족도] - 맛집에 대한 긍정/부정 표출, 만족도 계산됩니다. 3. SUSTAINABILITY [지속성] - 시기별 '맛집' 을 보여주는 기능은 없습니다. 4. COMPARE [비교] - 맛집 여러개를 한꺼번에 '비교' 하는 기능은 없습니다. 5. 정리 – 필터링, 만족도 가능하나, '맛집' 한정입니다.
	[소셜 매트릭스, PulseK] 1. FILTERING [필터링] - 따로 필터링 개념 기능은 제공하지 않습니다. 2. SATISFACTION [만족도] - 키워드에 따라 긍정/부정 추이를 알 수 있습니다. 3. SUSTAINABILITY [지속성] - 시기별 '컨텐츠' 의 긍정/부정 추이를 알 수 있습니다. 4. COMPARE [비교] - 컨텐츠를 한번에 '비교' 해주는 기능은 없습니다. 5. 정리 – 만족도, 지속성 부분 기능만 사용 가능합니다.
	[결론] 다이닝 코드는 '맛집' 이라는 한정된 컨텐츠를 가지고 있고, '맛집' 이라는 컨텐츠 특성 상 광고량에 대한 정보가 필요 없습니다. 또한 소셜매트릭스, PulseK는 연관 키워드, 긍정/부정에 대한 정보는 알려주지만 '광고 데이터' 에 대한 필터링이 없어 이게 가짜 리뷰를 많이 포함했다면 정확한 긍정, 부정 여부를 확인할 수 없습니다. 제가 만들 APP은 광고데이터와 긍정/부정을 화면에 명시적으로 표시하고, 랭킹을 사용자가 커스텀마이징 하여 데이터로 나타낼 수 있어 이 부분이 차별점이라 생각합니다

3. 어플리케이션 구조

그림 12, 13은 본 논문에서 설명하고 있는 ADDAVICH 시스템 및 구성을 보여주는 그림이다.

본 논문의 어플리케이션은 Android Studio와 PHP, MySQL, Tensorflow 및 형태소 분석 Python 어플리케이션을 사용하여 구성하였다.

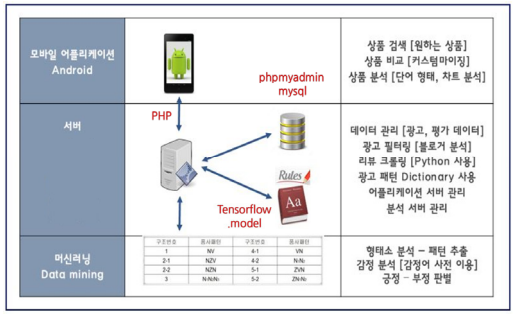


그림 12. 시스템 구성도 1
Fig. 12. System Configuration Plot 1

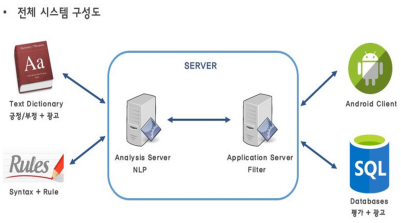


그림 13. 시스템 구성도 2
Fig. 13. System Configuration Plot 2

Ⅳ. 결론 및 향후 연구

본 논문에서는 개발한 ADDAVICHI 시스템에서 적용한 기술과 결과 시스템을 설명하였다.

홍보에만 치중하는 '바이럴 마케팅' 때문에 이용후기를 거짓으로 작성하거나 과장, 확대해 불신감을 안겨줄 수 있는 정보일지라도, 본 어플리케이션을 이용하여 필터링된 데이터를 보여줄 수 있어 사용자의 광고 노출을 줄일 수 있다. 사용자에게 좀 더 공정한 데이터를 제공하여 정보의 왜곡을 줄이고자 하였다.

현재는 광고 필터링에 머신러닝 기법으로 지속적으로 학습되고 누적되지 않아 이에 관한 추가적인 연구개발이 필요하다. 또한, GUI 디자인이 사용자의 관심을 끌기에 부족한 면이 있다. 실용화 단계에서는 사용자에게 보다 편리함을 줄 수 있는 기능을 추가할 필요가 있다.

현재는 네이버 블로그 정보만을 이용해서 필터링 하지만, 향후에는 정답 데이터를 학습시켜 좀 더 필터링 성능을 향상 시킬 계획이다. 표 4는 ADDAVICHI 시스템의 향후 개선 전략을 정리한 내용이다.

표 4. 단기, 중기, 장기 시스템 개선 전략
Table 4. System Enhancement Strategies

	단기 전략	중기 전략	장기 전략
대상	일반 사용자	일반 사용자 기업	일반 사용자 기업
비즈니스 전략	- 기술 커뮤니티 홍보 - 배너 광고 제외 - 무료 서비스 운영 - 페이스북 페이지 운영	- 상품별 비교 기능 추가 - WEB에 배너 광고 추가 - 기업 전용 심화 기능 (시간별 신뢰, 만족도)	- 통계 데이터 공개
제공 서비스	1차 : 런칭, 성능 테스트 2차 : 런칭, 모니터링 3차 : 기업 전용 분석 (APP 오픈)	1차 : WEB 서비스 오픈 2차 : 기업 전용 분석 (WEB, APP 오픈) 3차 : 심화 분석 레포트	1차 : 매체별 서비스 (리뷰, 뉴스, SNS) 2차 : 전 플랫폼 이용

References

[1] Viral Marketing, Wikipedia. Retreived at https://en.wikipedia.org/wiki/Viral_marketing on Aug. 15, 2019.
[2] "Advertising industry in 2016, 15 trillion 189.7 billion won," Newswatch. Retreived at http://www.newswatch.kr/news/articleView.html?idxno=13361 on Aug.15, 2019.
[3] J. Chang, A Study on Resesrch Trends of Graph-based Test Representations for Text Mining, JIIBC, Vol. 13, No. 5 | (2013) pp. 37~47.
[4] "[famous restaurant, uncomfortable truth]Unreliable online restaurant recommendation. "Customer deception" vs "information provision", heraldcorp, Retreived at http://news.heraldcorp.com/view.php?ud=20170617000024 on Aug. 15, 2019.
[5] Bidirectional recurrent neural networks. Retreived oat https://en.wikipedia.org/wiki/Bidirectional_recurrent_neural_networks1997. on Aug. 15, 2019.
[6] KoNLPy: Korean NLP in Python, KoNLPy.org. Retreived at http://KoNLPy.org/en/latest/ on Aug. 15, 2019.
[7] Beautiful Soup, Beautiful Soup Documentation. Retreived at https://www.crummy.com/software/BeautifulSoup/bs4/doc on Aug. 15, 2019.
[8] Word2vec, Wikipedia. Retreived at https://en.wikipedia.org/wiki/Word2vec on Aug. 15, 2019.
[9] Minsub Won,A Study on the Sentiment Analysis of Film Review Using BiLSTM, GitHub Project Documents, Retreived at https://github.com/MSWon/Sentimental-Analysis on Aug. 15, 2019. (In Korean)
[10] Softmax function, Wikipedia. Reteived at https://en.wikipedia.org/wiki/Softmax_function on Aug. 15, 2019.
[11] TensorFlow, Wikipedia. Retreived at https://en.wikipedia.org/wiki/TensorFlow on Aug. 15, 2019.
[12] Radim Řehůřek, Gensim, Wikipedia. Retreived at https://en.wikipedia.org/wiki/Gensim on Aug. 15, 2019.
[13] Introduction to Python's Collections Module, Stacj Abuse, Jan. 02, 2019. Retreived at https://stackabuse.com/introduction-to-pythons-collections-module/ on Aug. 15, 2019.
[14] Seleniu, SeleniumHQ. Retreived at https://www.seleniumhq.org on Aug. 15, 2019.
[15] OSS Project hanNanum, Semamtic Web Research Center. Retreived at http://semanticweb.kaist.ac.kr/hannanum/index.html

on Aug. 15, 2019.
[16] kkma Project, Seoul National University IDS Laboratory, Retreived at http://kkma.snu.ac.kr/ on Aug. 15, 2019.
[17] Yong Woon Lee, Young Ho Yoo, MeCab-Ko-Dic, A bunch of silver coins Project. Retreived at http://eunjeon.blogspot.com/ on Aug. 15, 2019.
[18] S. Baccianella, A Esuli, F Sebastiani, Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining., Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), May 2010.
[19] Web crawler, Wikipedia. Retreived at https://en.wikipedia.org/wiki/Web_crawler on Aug. 15, 2019.
[20] DiningCode,. Home Page. Retreived at https://www.diningcode.com on Aug. 15, 2019.
[21] Ga-ram Kim, Ung-mo Kim, "Utilization Plan of Graph-based Text Representation Model for Text Mining" Proceedings of KIIT Conference, 432-433, 2017.
[22] Eun-Sook Cho, Chul-Jin Kim, “A Customization Method for Mobile App.'s Performance Improvement” Journal of the Korea Academia-Industrial cooperation Society(JKAIS), Vol. 17, No. 11, pp. 208-213, 2016.

저 자 소 개

정 다 비 치(학생회원)

- 2013 ~ 현재: 한국산업기술대학교 컴퓨터공학과(4년 재학)
- 주 관심분야 : 빅데이타, 기계학습

노 영 주(정회원)

- 1984 : 고려대학교 공학사
- 1986 : FDU 전산학 석사
- 2000 : UNSW 컴퓨터공학 PhD
- 2005 ~ 현재 : 한국산업기술대학교 컴퓨터공학부 교수
- 관심분야 : SW, HCI, IoT, ML