**개별연구 결과 보고서**

사용자 리뷰 정보 기반

인위적 인기장소 판별 및 정보 제공 연구

|  |  |
| --- | --- |
| 과목명 | 개별연구(CS 웹정보의 빅데이터 분석을 통한 사용자 편의 서비스 연구) |
| 담당교수 | 정준호 |
| 학과 | 통계학과 |
| 학번 | 2017112282 |
| 이름 | 김현 |
| 제출일자 | 2022.07.13 |

개별연구 결과 요약

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1.인적사항 | | | | | | | |
| 소속 | | 동국대학교 이과대학 통계학과 | | | | | |
| 학번 | 2017112282 | | | 성명 | 김현 | 학년도/학기 | 2022학년도/여름학기 |
| 교과목명 | | | | 개별연구(CS 웹정보의 빅데이터 분석을 통한 사용자 편의 서비스 연구) | | 학수번호 | DES3912-01 |
| 2.개별연구 수행 결과 | | | | | | | |
| 연구목적  및  필요성 | | | 최근 인터넷 이용자들은 목적지를 정할 때 지도서비스나 sns등에서 미리 정보를 많이 찾는다. 이를 이용해 검색 서비스에서 상위에 뜨기 위하여 전문적인 리뷰 업체 등을 이용하는 사업장도 존재한다. 소비자의 입장에선 이런 업체와 실제로 소비자들에게 인정받아 상위에 뜬 업체들을 구분하여 올바른 정보를 받아들이는 것이 유익하다. 이에 개별연구의 주제에 맞게 장소 정보의 빅데이터를 활용하여 장소 정보 서비스 이용자들의 올바른 정보 습득의 편의를 향상시키기 위해 본 연구를 진행하였다. | | | | |
| 연구내용 | | | 본 연구에서는 사용자가 목적지를 검색할 때 가장 흔히 사용하고, 리뷰 정보가 가장 많이 저장돼 있는 네이버지도를 기준으로 데이터를 수집하고 사업장에 대한 분석을 진행하였다. 분석의 용이를 위해 검색창에 ‘강남 맛집’을 검색했을 때 나오는 장소 목록을 분석 범위로 잡았고, 각각의 사업장에 대해 블로그 리뷰의 여러 정보를 수집하고, 특정 문구와 로고를 분석하여 인공 리뷰를 판별한다. 또한 블로그의 전체 텍스트 내용에서도 인위적인 리뷰를 걸러내는 시도를 진행한다. | | | | |
| 연구결과 | | | 본 연구에서는 1차 목적으로 잡았던 블로그 리뷰 별 특정 문구, 특정 단어의 포함 여부를 통한 인위적 리뷰 판별에 성공하였다. 또한 2차 목적으로 잡았던 블로그 리뷰 별 광고 공시 로고를 통한 인위적 리뷰 판별에도 성공하였다. 3차 목표로 잡았던 블로그 리뷰의 전체 텍스트 정보를 통한 인위적 블로그 리뷰 판별 또한 시도했으나 유의미한 수준까지 끌어내려면 추가적인 자료 수집과 연구가 필요할 것으로 예상한다. | | | | |

사용자 리뷰 정보 기반

인위적 인기장소 판별 및 정보 제공 연구

Research on artificially popular spot identification and information provision based on user review information

요 약

가고자 하는 장소를 대부분 sns나 인터넷 지도 서비스로 찾아 결정하는 요즘, 해당 서비스들의 상위에 뜨는 장소들은 실제로 사용자들에게 인정받은 경우도 있지만, 전문 리뷰 업체 등을 이용해 인위적인 작업으로 만들어진 경우도 적지 않다. 이에 인위적으로 조작되지 않은 일명 ‘핫플레이스’에 대한 판별을 하고 이를 사용자에게 제공하는 발판을 만들고자 본 연구를 진행하였다. 연구 결과 특정 문구와 이미지내 문구를 기준으로 인위적인 리뷰를 판별 해내는데 유의미한 성과를 보였고, 실제로 지도 서비스 상의 ‘요즘 뜨는’ 순으로 상위에 뜨는 장소들이 인위적인 리뷰 작업을 많이 한 것으로 나타났다. 블로그 리뷰의 전체 텍스트 정보에서도 인위적인 리뷰들의 특징을 찾아내보려 했으나 추가적인 조사 및 연구가 필요할 것으로 판단된다.

1. 서론

인터넷으로 대부분의 정보를 찾고 접하는 요즘, 목적지를 정할 때도 sns 서비스나 지도 서비스를 이용하여 탐색부터 결정까지 하는 경우가 잦다. 특히 가볼 장소를 정할 땐 타인의 의견들을 참고하기 위해 지도 서비스를 이용하는 경우가 잦은데, 사업장들에 이 점을 이용하는 경우가 존재한다. 지도 서비스에서 상위에 뜨기 위해 인위적인 리뷰를 작성하거나, 전문 리뷰 광고 업체에 의뢰해 상위에 뜨게 조작을 하는 경우가 존재한다. 이에 따라 상위에 뜨는 업체들은 실제로 다수의 사용자들에게 인정받아 소위 말하는 ‘핫플레이스’가 된 곳과, 앞서 언급한 ‘인위적 핫플레이스’가 혼재하게 된다. 이는 사용자의 입장에서 잘못된 정보를 얻게 만들고, 설사 이 점을 인지하고 자연스레 인기 장소가 된 장소를 찾으려 해도 조작된 장소와 구분하기 쉽지 않다. 이 점에 착안하여 본 연구는 네이버 지도상에서 연구의 용이와 현실성을 고려해 ‘강남 맛집’을 고려했을 때, 기본적으로 선택되는 정렬 방식인 ‘요즘 뜨는’ 순으로 상위 업장들에 대한 블로그 리뷰를 크롤링한다. 해당 데이터로 인위적인 장소를 판단하고, 상위에 뜨는 장소들이 어떠한 형태를 보이는지 분석해보고자 한다. 이를 통해 지도 서비스의 상위에 뜨는 장소들이 실제로 인위적인 조작의 영향을 받았는지 확인하고, 일정 수준이상으로 인위적인 리뷰를 활용하여 상위에 뜬 장소들은 인위적 핫플레이스로 판단하여 지도 서비스 사용자에게 더욱 양질의 정보를 제공할 수 있는 기반을 마련할 수 있을 것이다.

해당 연구 주제와 관련하여 [1] Min Cheol Lee, Hyun Shik Yoon : A Study on Detecting Fake Reviews Using Machine Learning: Focusing on User Behavior Analysis 논문에서 진행한 연구가 존재한다. 해당 논문에서는 블로그 리뷰의 광고성을 확인하기 위하여 리뷰 자체의 내용에 더해 블로그 전체의 다양한 특성을 모두 변수로 사용하여 인위성 예측 테스트를 진행하였다. 본 연구는 이 연구를 참고하여 사용자에게 편의성을 준다는 점, 리뷰 자체적인 내용만으로도 인위성 예측 성능을 향상 시킨다는 점에 주안점을 두고 진행하였다.

본 보고서의 순서는 장소별 리뷰 데이터의 수집, 명백한 텍스트 근거를 통한 인위적 리뷰 판별, 광고 인증 로고를 이용한 인위적 리뷰 판별, 리뷰 텍스트 분석을 통한 인위적 리뷰 판별 시도, 향후 연구 순으로 진행된다.

2. 리뷰 데이터의 수집

리뷰 데이터는 네이버 지도의 검색창에 ‘강남 맛집’을 검색했을 때 ‘요즘 뜨는’ 순으로 상위 3페이지에 뜨는 장소를 기준으로 수집하였다. 각 장소들에서 블로그 리뷰를 최대 300개까지 크롤링 가능하게 설계한 후, html을 이용해 각 블로그의 순서 정보, 텍스트 내용, 이미지 내용, 작성 날짜 등을 수집하였다.

텍스트, 영수증, 스크린샷이(가) 표시된 사진

자동 생성된 설명

그림 : 수집한 데이터의 구조

그림1은 ‘강남 맛집’을 네이버지도 검색창에 검색 했을 때 상위에 뜨는 장소 중, 총 113개의 장소에서 총 5720개의 블로그에 대한 텍스트 정보를 끌어온 데이터프레임의 구조도이다. 그림1과 같이 장소와, 블로그, 작성 날짜, 텍스트 내용, 특정 문구 포함 여부에 따른 인위적 리뷰 여부가 place, blog, date, content, altificial\_simple 이란 이름으로 변수로 포함 돼있다.

3. 명백한 텍스트 근거를 통한 인위적 리뷰 판별

블로거를 통한 블로그 광고 시, 업체로부터 일정의 대가를 받고 작성한 것이라는 표시를 해야하는 법적 의무가 있다. 따라서 이를 잘 지킨 블로그들은 일반적으로 텍스트상에 ‘본 포스팅은 업체로부터 소정의 원고료를 받고 작성된 글입니다’ 같이 명확하게 금전적 대가가 오간 글인 것을 알 수 있게 표시를 한다. 이 문장의 구조는 블로그마다 제각기 다른데, 본 연구에서는 공통적으로 자주 등장하는 광고 인증 문구 내 단어인 ‘원고료’, ‘고료’, ‘소정의’, ‘업체로부터’, ‘쿠팡파트너스’, ‘제공’, ‘업체로부터’, ‘금전적’ 들이 or 조건으로 포함되면 광고 리뷰로 분류하였다. 그 결과 텍스트 데이터가 수집된 총 5720개의 블로그 중 652개의 블로그가 인위적 블로그 리뷰로 분류되었다.

4. 광고 인증 로고를 이용한 인위적 리뷰 판별

금전적 대가가 오고 갔다는 표시를 그림2처럼 로고를 통해 알리는 경우도 존재한다.

텍스트, 포유류이(가) 표시된 사진

자동 생성된 설명

그림 : 금전적 대가 표시를 한 블로그 내 로고

기존에 이런 로고와 일치하는 이미지가 있는지 여부로 로고를 통해 인위적인 블로그 리뷰들을 분류해보고자 하였으나, 해당하는 로고의 종류가 너무 다양하다고 판단하여 이미지 내의 텍스트를 인식해서 3번단계에서와 같은 문구들이 있는지 여부로 인위적인 블로그를 선별하기로 하였다. 구글의 Tesseract 라이브러리를 사용하여 텍스트 인식에 사용하였고, 그 결과 이미지 수집이 가능했던 5021개의 블로그 중 18개의 블로그에서 광고 리뷰가 확실하다고 판단될 만한 로고를 탐지하였다. 해당되는 블로그 중 10개의 블로그가 3번 단계에서 인위적인 블로그로 분류되지 않아 추가로 분류해준 결과 662개의 블로그 리뷰가 인위적인 리뷰로 분류되었다.

이후 이렇게 분류된 블로그 리뷰 정보들을 이용해 각 업장마다 몇 개의 인위적인 리뷰를 사용했는지 분석하고 상위에 뜬 장소순으로 인위적 리뷰 사용량을 그림3과 같이 시각화 해보았다.

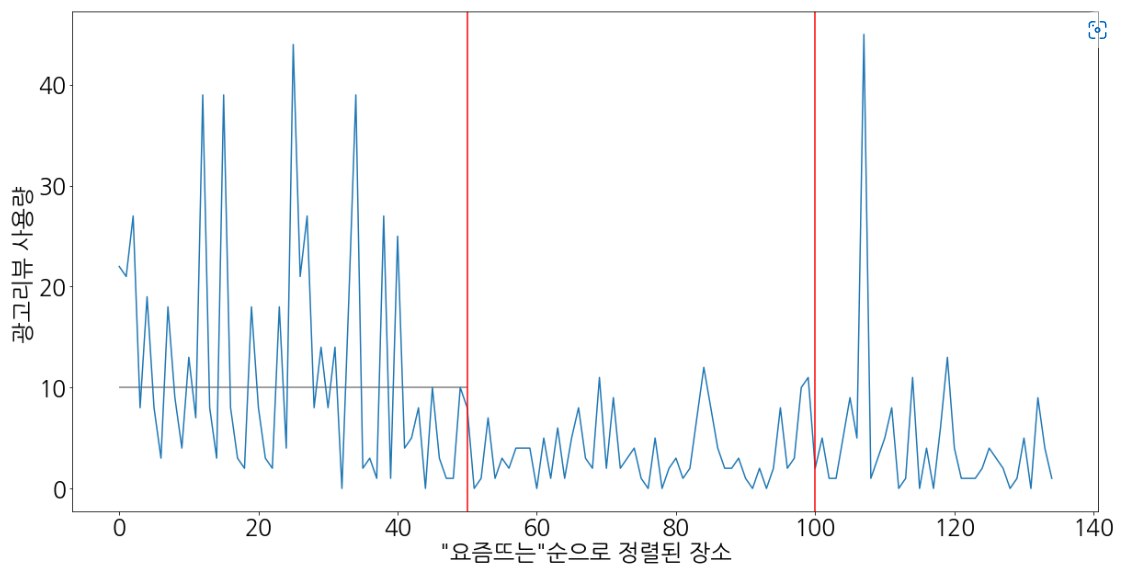


그림 : 상위에 뜬 블로그 순서대로 사용된 광고리뷰 시각화 그래프

그림3에서 볼 수 있듯이 상위에 분포한 장소들이 광고리뷰로 분류된 블로그 리뷰를 많이 포함하는 것을 알 수 있다. 즉 검색 시 앞에 떠서 사용자가 접하기 쉬운 장소일수록 인위적으로 끌어올려진 장소일 가능성이 높은 것이다. 이는 본 연구의 필요성을 입증하고, 자연스럽게 상위에 뜬 장소들과 그렇지 않은 장소들을 구분하는 것이 유의미하다는 점을 보여준다. 또한 앞쪽에 분포한 장소들 중 인위적 리뷰를 거의 사용하지 않은 업장도 존재하기 때문에 해당 업장들이 공정한 평가를 받는데 본 연구가 진행한 라벨링이 도움을 줄 수 있다. 또 한가지 눈여겨 본 점은 광고리뷰를 많이 사용한 장소와 그렇지 않은 장소가 대부분 번갈아서 등장하는 양상을 보였다는 점이다. 지도 서비스에서 보이지 않는 부분의 설계를 유추해볼 수 있는 부분이었다.

5. 리뷰 텍스트 분석을 통한 인위적 리뷰 판별 시도

본 연구와 비슷한 목적을 가지고 진행한 연구인 ‘머신러닝을 활용한 가짜리뷰 탐지 연구: 사용자 행동 분석을 중심으로’(이민철, 윤현식)에서는 블로그 리뷰의 인위성을 판별하는데 있어 본연구가 수집한 정보 외에도 해당 블로거와 블로그 자체에 대한 정보도 폭넓게 수집하였다. 하지만 해당 연구에선 텍스트 자체적인 내용에 대한 분석을 통해 가짜리뷰를 탐지하는 부분은 자세히 다루지 않았다. 본 연구는 분석과 탐지의 효율성을 증가시키고자, 텍스트 정보만으로도 숨겨진 광고 리뷰를 끌어 내어 유의미한 결과를 도출하고자 하였다.

리뷰의 구조적인 특징들도 변수로 사용할 시 유의할 수 있다는 가정을 세웠기 때문에 다음과 같은 예측변수(독립변수)들을 생성해주었다. 독립 변수로 사용하기 위해 크롤링된 텍스트의 전체 길이를 뜻하는 content\_len, 특수기호와 html코드가 제거된 후 주요 형태소만 남긴 텍스트 전체 길이를 뜻하는 content\_new\_len, 둘 사이의 차이를 뜻하는 surplus\_text, 사용된 이미지 개수를 뜻하는 image\_count를 변수로 만들었다. 또한 리뷰의 내용으로도 인위적인 리뷰들을 어느정도 판별하고 예측 성능을 향상 시킬지 알아보기 위하여 각 블로그의 전처리 된 문장을 워드임베딩하여 변수로 만들어 주었다. 워드임베딩을 하는데 있어 추후 연구에 적용될 때의 일반화를 위해 stopwords는 최소한으로 적용하고 100차원의 워드벡터 공간을 그림4와 같이 만들어주었다.

테이블이(가) 표시된 사진

자동 생성된 설명

그림 : 100차원 워드벡터 공간

여기에 앞선 단계에서 만들어 놓았던 리뷰 구조적 특징을 나타내는 변수들을 다시 추가한 후 각 단어들과 해당 변수들을 독립변수로 삼는 분류 모델을 적합 해보았다.

분류 모델은 로지스틱 회귀 모델과 랜덤포레스트 분류 모델을 사용하였다. 그 결과 Full-model을 기준으로 로지스틱 회귀모델은 그림5의 수정된 R제곱값과 같이 분류를 제대로 해내지 못하는 모습이었고(1에 가까울 수록 좋은 예측 성능을 뜻한다),

테이블이(가) 표시된 사진

자동 생성된 설명

그림 : 로지스틱 회귀모델 적합

랜덤포레스트 모델은 그림5와 같이 76%정도의 정확도를 보이며 앞선 모델보다 훨씬 준수한 분류 성능을 보여주었다.

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 : 랜덤포레스트 모델 적합 결과

이후 적합한 랜덤포레스트 모델에서 110개의 전체 변수 중 어떤 변수들의 중요도가 높은 지 산출해 보았다. 그 결과 그림7과 같은 결과를 확인하였는데,

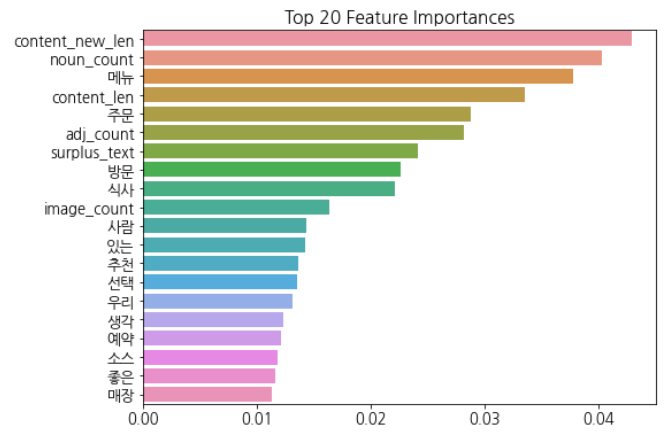


그림 : 최종 랜덤포레스트 모델의 내림차순 변수 중요도 top20 그래프

html코드와 이모티콘을 제거하고 형태소 분석을 완료한 이후의 리뷰 길이가 가장 주효한 변수인 것으로 나타났고, 그에 따라 문장이 길어질수록 당연히 그 크기가 늘어날 수밖에 없는 noun\_count가 그 다음 순위의 중요도를 나타내었다. ‘메뉴’, ‘주문’, ‘방문’, ‘식사’ 등의 감정이 배제된 형식적인 언어들이 예측 중요도 상위에 포함되었다. 또한 사용된 형용사의 수를 뜻하는 adj\_count도 상위권을 차지한 것으로 보아 형용사가 비교적 많이 사용됐을 때도 인위적인 리뷰일 가능성이 높아진다는 것을 확인할 수 있었다.

리뷰 자체의 내용에 집중하여 전체적인 블로그나 블로거에 대한 수집은 진행하지 않은 탓에 탓에 높은 수준의 성능은 보여주지 못하였다 판단한다. 하지만 랜덤 포레스트 모델 적합 결과 리뷰 내용만의 구조적 특징과 포함 단어만으로도 어느 정도는 분류해 낼 수 있다는 사실을 알 수 있다.

6.결론 및 향후 연구

본 연구에서는 인위적인 광고 리뷰를 이용하여 지도서비스 상위에 뜨는 업체들을 분류해내고, 더 나아가 분류 성능을 향상시킬 수 있는 방법에 대해 분석을 진행하였다.

분석 결과 수집된 데이터를 기준으로 인위적인 광고 리뷰일 경우, 특정 문구로 광고 리뷰임을 알리는 경우가 대부분이었고, 광고임을 나타내는 로고로 알리는 경우는 예상보다 많지 않았다.

상기 방법들을 통해 분류해낸 블로그들을 기준으로, 수집한 각 장소들이 얼마나 많은 강고리뷰를 사용하였는지 할당하고 분석해보았다. 예상대로 앞 순서에 뜨는 장소들 중에 광고리뷰를 사용한 장소들이 상당수였고, 이를 통해 본 연구가 가지는 첫번째 목적인 ‘인위적인 인기 장소의 분류를 통한 사용자 편의성 향상’이 정당한 필요성을 가진다고 확인할 수 있었다. 또한 장소마다 낸 광고 리뷰 총 사용 수 통계를 통해, 상위에 뜬 장소들과 그렇지 않은 장소들이 번갈아 나타나게 설계되어 있단 점도 눈여겨볼 부분이었다.

위와 같은 단계들을 통해 기본적으로 인위적인 인기 장소들을 분류해낸 후, 그 분류의 성능을 향상시키기 위한 방법을 시도하였다. 블로그 텍스트 전체 내용에서 뽑아낼 수 있는 정보들을 가지고 예측 분류 모델의 성능을 향상시키기 위한 여러 방식을 시도한 결과, 글의 내용적인 측면에서는 몇 가지 제약점에 의해 유의미한 예측 성능 향상을 시키지 못하였지만, 블로그 글의 구조적인 여러가지 특성들은 인위적 블로그 글의 분류에 유의미한 성능향상을 일으킨다고 판단할 수 있었다.

향후에 더 방대한 수준의 데이터 수집과 각 블로거에 대한 평가도 변수로 사용하여 성능향상을 목표로 연구를 진행한다면, 더 나은 인위적 인기장소 판별 모델을 만들어 낼 수 있을 것이다.

참고문헌

[1] Min Cheol Lee, Hyun Shik Yoon : A Study on Detecting Fake Reviews Using Machine Learning: Focusing on User Behavior Analysis

[2] [GitHub - tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (main repository)](https://github.com/tesseract-ocr/tesseract)