JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 8, pp. 39-46, August 2021
https://doi.org/10.9708/jksci.2021.26.08.039

Social Big Data Analysis for Franchise Stores

Hyeon Gyu Kim*

*Associate Professor, Div. of Computer Science and Engineering, Sahmyook University, Seoul, Korea

[Abstract]

When conducting social big data analysis for franchise stores, reviews of multiple branches of a franchise can be collected together, from which analysis results can be distorted significantly. To improve its accuracy, it should be possible to filter reviews of other branches properly which are not subject to the analysis. This paper presents a method for social big data analysis which reflects characteristics of franchise stores. The proposed method consists of search key configuration and review filtering. For the former, the open data provided by Small Business Promotion Agency is used to extract region names for collecting reviews more accurately. For the latter, open search APIs provided by Naver or Kakao are used to obtain franchise branch information for filtering reviews of other branches that are not subject to analysis. To verify performance of the proposed method, experiments were conducted based on real social reviews collected from online, where the results showed that the accuracy of the proposed review filtering was 93.6% on the average.

▸ **Key words**: Big data analysis, Social reviews, Franchise analysis, Open data, Review filtering

[요 약]

프랜차이즈 스토어를 대상으로 소셜 빅데이터 분석을 수행할 경우, 프랜차이즈에 속한 여러 분점의 리뷰들이 함께 수집될 수 있어 분석 결과가 왜곡될 수 있다. 이 경우 분석 정확도를 높이기 위해서는 분석 대상이 아닌 타 분점의 리뷰들을 적절히 필터링할 수 있어야 한다. 본 논문에서는 프랜차이즈 스토어들의 특성을 반영한 소셜 빅데이터 분석 방법을 제안한다. 제안 방법은 검색어 설정 방법과 리뷰 필터링 방법을 포함한다. 검색어 설정을 위해, 소상공인진흥공단에서 제공하는 공공데이터를 기반으로 검색에 필요한 지역명을 추출한다. 그리고 리뷰 필터링을 위해, 네이버 및 카카오 등에서 제공하는 검색 API를 이용하여 프랜차이즈 분점 정보를 알아내고, 분석 대상이 아닌 타 분점의 리뷰들을 필터링하는데 이용한다. 제안 방법의 검증을 위해 온라인에서 수집된 실제 리뷰를 대상으로 실험을 수행하였으며, 제안 방법의 리뷰 필터링 정확도는 평균 93.6%로 조사되었다.

▸ **주제어**: 빅데이터 분석, 소셜 리뷰, 프랜차이즈 분석, 공공데이터, 리뷰 필터링

• First Author: Hyeon Gyu Kim, Corresponding Author: Hyeon Gyu Kim
*Hyeon Gyu Kim (hgkim@syu.ac.kr), Div. of Computer Science and Engineering, Sahmyook University
• Received: 2021. 06. 10, Revised: 2021. 07. 20, Accepted: 2021. 07. 27.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

SNS 피드 및 블로그 리뷰 등을 포함한 소셜 빅데이터는 고객 관점의 의견이나 불만 사항을 추출하기 위한 목적으로 많은 응용에서 활용되고 있다[1, 2]. 소셜 빅데이터는 네이버 및 구글 등의 온라인 포털 업체에서 제공하는 오픈 API[3, 4]를 통해 무료로 획득 가능하며, 신용카드, 휴대폰 이용 내역 등 수치로 이루어진 다른 형태의 빅데이터에 비해 고객들의 의견이나 불만 사항을 텍스트 형식으로 바로 파악할 수 있다는 점에서 선호되고 있다.

소셜 빅데이터 분석은 텍스트를 다뤄야 한다는 점에서 세밀한 주의가 필요하다. 특히 검색어를 어떻게 설정하느냐에 따라 분석 정확도가 크게 달라질 수 있다. 너무 일반적인 검색어를 이용할 경우, 온라인으로부터 수집된 리뷰에는 연관성이 다소 떨어지는 노이즈 리뷰가 다수 포함되어 분석 정확도가 떨어질 수 있다. 이에 반해 검색어를 너무 세분화할 경우에는 오히려 주어진 검색어를 만족하는 리뷰 수가 적어져 정확도가 떨어지는 문제가 발생한다.

이러한 문제점은 주어진 분석 대상이 프랜차이즈 스토어일 경우에도 동일하게 나타난다. 예를 들어, 울산에 위치한 "구구양꼬치"에 대한 키워드 분석을 수행한다고 가정해보자. "스타벅스"나 "맥도날드" 등 대중들에게 알려진 프랜차이즈가 아닐 경우, 주어진 상호명의 프랜차이즈 여부를 인지하지 못하고 분석을 수행할 수 있다. "구구양꼬치"의 경우, 울산을 포함한 부산, 창원 지역에 4개의 매장을 지닌 프랜차이즈 스토어이다. 따라서 검색어를 단순히 "구구양꼬치"로 설정할 경우, 타 지역 매장의 리뷰까지 모두 합산하여 분석을 수행하게 되므로, 왜곡된 분석 결과를 얻게 된다.

타 지역 매장의 리뷰를 제외시키기 위해 검색어를 보다 세밀하게 설정할 수 있다. 예를 들어 "구구양꼬치"가 위치한 지역명을 검색어에 포함시켜, "구구양꼬치 울산 신정동"으로 검색어를 설정할 수 있다. 이 경우 타 지역 매장의 리뷰는 제거할 수 있으나, 정작 울산 매장을 언급하고 있는 진성 리뷰들이 다수 제외될 수 있다. 이는 다수의 리뷰가 행정 지역명(신정동)이 아닌 "구구양꼬치 삼산" 등 대중들이 익숙한 지역명으로 스토어를 언급하기 때문이다. (홍대, 서면 등과 같이, 삼산은 신정동과 달동을 포함한 울산 남구의 상권을 대중적으로 지칭하는 지역명이다.)

검색 키워드를 "구구양꼬치 삼산"으로 설정하는 것 역시 답이 될 수는 없다. 삼산 지역 내에 여러 프랜차이즈 스토어가 존재할 수 있기 때문이다. 예를 들어, 삼산 지역에 포함된 신정동과 달동에 "구구양꼬치" 분점이 각각 위치하고 있을 경우, 위와 같은 키워드를 이용한다면 여전히 리뷰가 합산되어 분석이 수행되는 문제점이 있다. 결국 위 문제를 해결하기 위해서는 각 스토어가 주로 어떤 지역명과 함께 언급되는지 파악해야 하며, 지역별 분점 정보도 함께 고려하여 검색어를 설정하고 리뷰 분석을 수행해야 함을 알 수 있다.

본 논문에서는 프랜차이즈 스토어를 대상으로 소셜 빅데이터 분석을 수행할 때 분석 정확도를 높이기 위한 방법을 제안한다. 위 예제에서 살펴본 바와 같이, 분석 대상이 아닌 타 분점들의 리뷰를 정확히 걸러내려면 검색어에 적절한 지역명이 함께 포함되어야 한다. 따라서 제안 방법에서는 검색 키워드 구성을 위해 소상공인진흥공단에서 제공하는 공공데이터를 이용하여 지역명을 추출하는 방법을 소개한다.

이에 반해 공공데이터에서 일부 스토어 정보가 누락될 경우, 프랜차이즈 지점들의 지역명이 명확히 추출되지 않을 수 있다. 온라인에서 수집된 실제 데이터를 대상으로 실험한 결과를 살펴보면, 공공데이터 누락으로 인해 발생하는 오류의 비율이 평균 34% 정도로 상당한 비중을 차지함을 알 수 있었다. 이러한 문제점을 보완하기 위해 네이버나 카카오 등에서 제공하는 검색 API를 활용하여 지역별 프랜차이즈 분점 정보를 알아내고, 이를 기반으로 분석 대상이 아닌 타 지역 분점들의 리뷰를 걸러내기 위한 방법을 함께 소개한다.

II. Related Work

소셜 빅데이터는 신용카드, 휴대폰 이용 내역 등 수치로 이루어진 다른 형태의 빅데이터에 비해 고객들의 의견이나 불만 사항을 텍스트 형식으로 바로 파악할 수 있다는 측면으로부터, 다양한 연구 주제에서 활용되고 있다. 예를 들어, 소셜 리뷰로부터 잠재적인 광고 키워드를 추출하거나[5], 블로그 리뷰로부터 영화 흥행 요인 분석[6, 7], 온라인 쇼핑몰의 상품평 분류를 통한 감성 분석 [8] 등의 주제가 소개된 바 있다.

특히 최근에는 기계학습 모델을 이용한 감성 분석 연구가 활발하게 진행되고 있다. 예를 들어 [9]에서는 순환 신경망(Recurrent Neural Network) 기반의 Variational Inference 모델을 제안하고 영화 리뷰 데이터 셋에 적용하여 높은 분류 정확도를 달성하였다. [10]에서는 LSTM (Long Short-Term Memory)을 기반으로 한 Parallel Stacked Bidirectional LSTM 모델을 제안하고 리뷰 분류에 적용하였다. 마찬가지로 [11-13]에서 순환 신경망이나 LSTM 알고리즘을 활용하여 영화 리뷰를 대상으로 감성 분석을 수행하였다.

소셜 빅데이터 분석 기술과 관련해서도 다양한 주제가 논의되었다. [1]은 소상공인들의 마케팅을 돕기 위한 서비스 지표들을 추출하고 이를 구현한 시스템을 소개하였다. 그리고 소셜 빅데이터 분석에 있어 형태소 분석 및 노이즈 리뷰 필터링의 중요성을 언급하였다. [14, 15]는 소셜 빅데이터 분석의 특성상 신조어나 고유 명사가 많아, 사전에 기반한 기존의 형태소 분석기법 등을 활용할 경우 정확도가 저하될 수 있음을 지적하였다. 그리고 문제점 해결을 위해 브랜칭 엔트로피[16, 17] 개념이나 비지도 기계학습법을 활용한 형태소 분석 방법[18] 등이 효과적으로 이용될 수 있음을 제안하였다.

노이즈 리뷰 필터링과 관련하여, [1]에서는 패턴 매칭을 이용하여 주어진 키워드와 관계없는 노이즈 리뷰를 추출하고 제거하는 기법을 소개하였다. 이에 반해, [19]에서는 여행지 평점 예측을 위해 기계학습 알고리즘인 합성곱 신경망(Convolutional Neural Network) 기반의 필터링 방법을 제안하였으며, [20]에서는 협업 필터링(Collaborative Filtering)을 기반으로 리뷰의 연관도를 수치화하고 필터링하는 방법을 논의하였다. 현재까지 논의된 알고리즘 중, 프랜차이즈 스토어를 대상으로 소셜 빅데이터 분석의 정확도를 높이기 위한 관련 연구는 찾기 어려운 실정이다.

위에서 소개한 참조 문헌들은 소셜 빅데이터 분석 단계에 따라 아래 표와 같이 분류될 수 있다. 소셜 빅데이터 분석 단계는 [그림 1]과 같이 도식화될 수 있으며, 아래에서 설명한다.

Table 1. Classification of References by social big data analysis steps

Analysis Step	Techniques	Reference #
Review Collection	Open search API, Web crawling	[1, 14]
Morphological Analysis	Statistical approach	[14-18]
	Dictionary-based approach	[21, 22]
Noise Review Filtering	Pattern matching	[1]
	Machine learning	[19, 20]
Review Analysis and Applications	Sentiment analysis	[8-13]
	Keyword analysis	[5, 14, 15]
	Miscellaneous	[6, 7]

III. Proposed Method

1. Overview

소셜 빅데이터 분석 과정은 크게 소셜 리뷰 수집, 형태소 분석, 노이즈 리뷰 필터링, 리뷰 분석, 시각화 등의 과정으로 구성된다(그림 1). 소셜 리뷰 수집 단계는 주어진 검색어를 만족하는 SNS 피드 글이나 블로그 리뷰 등을 온라인 사이트로부터 수집하는 과정이며, 정확한 분석 결과를 얻기 위해서는 해당 단계에서 검색어를 정확하게 설정하는 것이 무엇보다 중요하다.

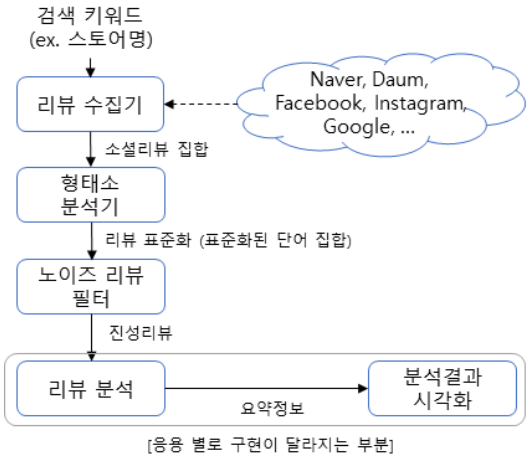

Fig. 1. Overview of social big data analysis

형태소 분석 단계는 수집된 리뷰에 대한 표준화된 단어 집합을 얻는 과정이다. 예를 들어, 리뷰에 포함된 형용사나 동사에 대해서는 다양한 활용 형태에 대한 원형 단어로 변환하는 과정이 필요하며, 분석에 영향을 미치지 않는 부사나 조사는 제거될 수 있다. 특히 소셜 리뷰는 신조어나 고유 명사를 포함하는 경우가 많으며, 꼬꼬마[21]나 한나눔[22] 등의 사전에 기반한 기존 형태소 분석기로는 정확한 형태소 분석 결과를 얻기 어려운 경우가 많다. (사전에는 신조어나 고유명사가 포함되어 있지 않기 때문이다.) 이러한 경우, 단어 빈도수 등의 통계 데이터에 기반한 형태소 분석 기법[15-18]들이 효과적으로 이용될 수 있다.

노이즈 리뷰 필터링 단계는 주어진 검색 키워드와 관계가 없는 노이즈 리뷰들을 걸러내는 단계이다. 일반적으로 소셜 리뷰 수집 단계를 통해 얻은 리뷰는 주어진 검색어가 포함되기만 하면 관련 있는 리뷰로 간주하고 검색 결과에 포함시킨다. 그러나 전달된 결과를 자세히 살펴보면 검색어와 관계가 먼 소위 "노이즈" 리뷰들이 50% 이상 포함된다. 예를 들어, 검색어로 "구구양꼬치"가 주어질 경우, "구구양꼬치" 근처의 학원이나 해당 건물 2층의 미용실 등, 주어진 검색어가 포함되지만 실제는 다른 스토어를 지칭하는 리뷰들이 다수 포함된다. 따라서 분석 정확도를 높이기 위해서는 이들을 필터링해야 하며, 이 과정에서 회귀분석이나 딥러닝 기법 등이 활용될 수 있다.

리뷰 분석 및 시각화 단계는 응용 영역에 따라 달라진다. 예를 들어 키워드 분석일 경우, 노이즈 필터링 단계를

통해 얻은 진성 리뷰들의 단어들을 대상으로 빈도수가 높은 단어들을 추출하여 키워드 분석 결과로 제공할 수 있으며, 시각화를 위해 jQCloud[23] 등의 다양한 키워드 클라우드 솔루션을 활용할 수 있다. 이 외에 감성 분석이나 온라인 평판 분석 등, 다양한 응용 목적에 맞게 리뷰를 분석하거나 시각화할 수 있다.

프랜차이즈 스토어를 대상으로 소셜 빅데이터 분석을 수행할 경우, [그림 1]의 과정에서 두 부분이 추가되어야 한다. 첫째는 소셜 리뷰 수집 단계에서 검색어를 설정하는 부분이며, 스토어가 위치한 지역명을 추가하여 검색어를 세분화하는 과정이 필요하다. 둘째는 노이즈 리뷰 필터링 단계이며, 회귀분석이나 딥러닝 등을 이용하여 필터링을 수행한 후 남은 진성 리뷰를 대상으로, 프랜차이즈 매장 정보를 이용하여 타 지역 매장을 언급하는 리뷰를 걸러내기 위한 추가적인 필터링 과정을 수행해야 한다. 각각의 과정에 대해 아래에서 설명한다.

2. Configuring Search Keywords

스토어에 대한 정보는 소상공인시장 진흥공단[24]에서 발행한 공공데이터로부터 얻어올 수 있다. 해당 데이터는 스토어 정보를 광역별(광역시나 도 단위)로 모아 엑셀 파일 형태로 제공되며, 스토어별로 상호명, 주소(행정동명 포함), 업종 등의 정보를 포함한다. 문제는 프랜차이즈 여부가 불분명하게 표기되어 있다는 점이다. 따라서 공공데이터의 상호명을 기반으로 소셜 리뷰 분석을 그대로 수행하게 될 경우, 1장에서 언급한 여러 문제가 발생할 수 있다.

각 스토어에 대해 프랜차이즈 여부를 알아내기 위해서는 다른 지역 데이터에서 동명의 스토어가 존재하는지 파악해야 한다. 예를 들어, "구구양꼬치"가 울산의 공공데이터 이외에 부산 및 경남 지역의 데이터에도 나타나는지 확인한다. 타 지역 데이터에 이름이 존재한다면, 해당 스토어는 프랜차이즈 스토어일 확률이 높다. 이로부터 검색어를 구성할 때 상호명 이외에 지역명을 함께 추가하여 검색어를 생성하고 소셜 리뷰 수집 단계를 수행한다.

검색어에 포함될 지역명은 광역 이름이나 시군구명, 또는 행정동명 중 하나 이상의 조합이 될 수 있다. 먼저 상호명이 광역 데이터에서 유일하게 나타나는 경우, 해당 스토어가 광역시 단위에 속해 있다면 지역명을 광역시명으로 설정한다. 이에 반해 스토어가 도 단위에 속해 있다면, 스토어가 속한 시군구명을 지역명으로 설정한다. 예를 들어, "구구양꼬치"가 울산광역시에 유일하게 나타난다면 지역명은 "울산"이 되며, "구구양꼬치 울산"을 검색어로 설정한다. 그렇지 않고 해당 스토어가 경상남도 데이터에서 유일하게 나타나며 창원시에 속해 있다면, 지역명은 "창원"이 되어 "구구양꼬치 창원"이 검색어로 설정된다. 도 이름을 지역명으로 설정하지 않는 이유는 대부분의 리뷰에서 도 단위의 이름보다 시군구 이름으로 스토어를 언급하는 경우가 많기 때문이다.

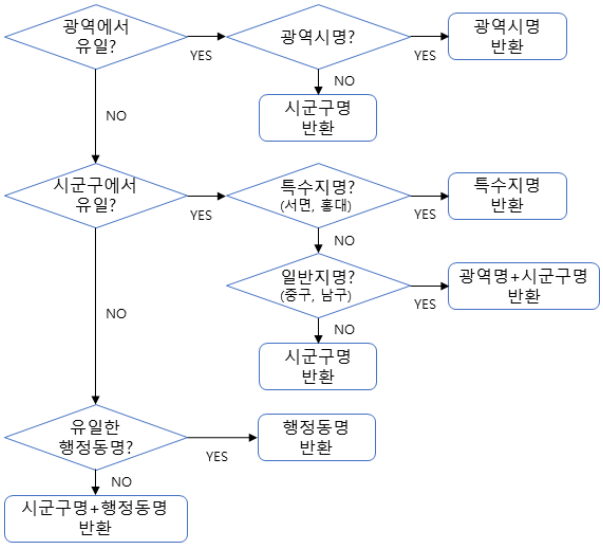

Fig. 2. Algorithm for region name selection for configuring search keywords

상호명이 하나의 광역 데이터 내에서 여러 번 나타날 경우, 시군구 단위로 더욱 세분화하여 상호명 출현 빈도를 체크한다. 만약 시군구 단위 내에서 상호명이 유일하다면, 해당 시군구명을 지역명으로 설정한다. 단 광역시의 경우, "동구", "서구", "남구", "북구", "중구" 등의 일반적인 이름이 시군구 단위 이름으로 설정될 수 있다. 이 경우 해당 이름만으로는 타 지역과 구분이 불가능하므로 광역 이름을 함께 써준다. 예를 들어 "구구양꼬치"가 울산 남구에 위치해 있다면, 검색어는 "구구양꼬치 남구"가 아닌 "구구양꼬치 울산 남구"로 설정된다.

위 내용에 대한 예외적인 상황으로, 만약 울산 남구의 대중적인 상권을 지칭하는 "삼산" 등의 단어가 별도로 존재한다면, 해당 지역명을 검색어로 이용한다. 이 경우 "구구양꼬치 울산 남구"가 아닌, "구구양꼬치 삼산"이 검색어로 설정된다. 이는 "부산진구"라는 이름 대신 "서면"이라는 이름이 더욱 대중적으로 이용되는 것과 동일한 맥락이다. 단, 해당 스토어가 "삼산"이나 "서면"이라는 지역에 속한 행정동명에 위치해 있는지 추가적인 체크가 필요하다.

만약 상호명이 특정 시군구 내에서 여러 번 나타난다면, 스토어가 위치한 행정동명을 지역명으로 설정한다. 예를 들어 "구구양꼬치" 분점이 울산 남구의 신정동에도 있고 달동

에도 위치해 있다면, "구구양꼬치 신정"이나 "구구양꼬치 달동" 등 행정동명을 이용해 검색어를 더욱 세분화한다. 단, 일부 행정동명의 경우 타 지역에 동일한 이름이 있을 수 있다. 예를 들어 "신정동"의 경우, 울산 남구 외에 서울 마포구에도 동일한 행정동명이 존재한다. 이러한 경우 해당 동명 외에 시군구 이름을 추가하여 지역명이 결정되며, 위 예제의 경우 "구구양꼬치 울산 신정" 등이 검색어로 채택된다.

3. Filtering Franchise Noise Reviews

검색어에 지역명을 정확하게 추가하려면, 공공데이터가 정확해야 한다는 가정이 필요하다. 그러나 이러한 가정은 100% 성립되지는 않는다. 이는 공공데이터의 특성 상 업데이트가 빠르지 않다는 점에 기인하며, 이로부터 최근 핫플레이스로 부상하는 신규 프랜차이즈 스토어들이 누락될 수 있다.

실제 울산 지역의 공공데이터를 대상으로 실험한 결과, 본점 정보만 등록되고 분점 정보가 등록되지 않아 프랜차이즈로 인식되지 않는 경우가 다수 발생하였다. 그리고 이들로부터 발생하는 노이즈 리뷰의 양이 전체의 34% 정도를 차지하는 것으로 조사되었다(4장 참고). 따라서 만족할 만한 분석 정확도를 얻기 위해서는 공공데이터로부터 알아낼 수 없는 프랜차이즈 분점 정보를 알아내고 관련 리뷰들을 삭제하기 위한 추가적인 방법이 필요하다.

제안하는 방법에서는 문제 해결을 위해 네이버나 카카오 등에서 제공하는 검색 API를 추가적으로 이용한다. 검색 API 호출을 최소화하기 위해, 먼저 [그림 1]의 노이즈 리뷰 필터링 단계 이후 얻어진 진성 리뷰들을 대상으로 키워드 추출을 수행한다. 그리고 추출된 키워드 중 스토어가 위치한 지역이 아닌 타 지역명이나 행정동명 관련 키워드의 빈도수가 많을 경우, 프랜차이즈 스토어로 간주하고 검색 API를 호출한다. 이러한 선별 과정이 필요한 이유는 공공데이터에 등록된 스토어의 수가 많기 때문이다. (울산에만 53,000여 개의 스토어가 등록되어 있다.) 따라서 프랜차이즈 여부를 별도로 판단하여 분석 효율을 높일 필요가 있다.

검색 API 호출 결과, 2개 이상의 스토어 정보가 반환되면 프랜차이즈 스토어로 간주한다. 이 경우, 반환된 스토어로부터 지역 및 행정동명을 추출하여 분점 위치 정보를 구성한다. 그리고 분석에 이용되었던 기존 리뷰 중 분점 지역명이 들어간 리뷰들을 노이즈로 간주하고 삭제 처리한다. 마찬가지로 3.2절에서 설명한 방법에 따라 상호명을 업데이트한다.

위 내용을 반영한 소셜 빅데이터 분석 과정은 아래 그림과 같이 도식화될 수 있다. [그림 1]과 비교하였을 때, 소셜 리뷰 수집 단계 이전에 공공데이터를 기반으로 지역명을 포함한 검색어 구성 단계가 새로이 추가되었으며, 노이즈 리뷰 필터링과 리뷰분석 단계 사이에 위에서 설명한 타 지역 분점 관련 리뷰를 삭제하는 단계가 추가되었다.

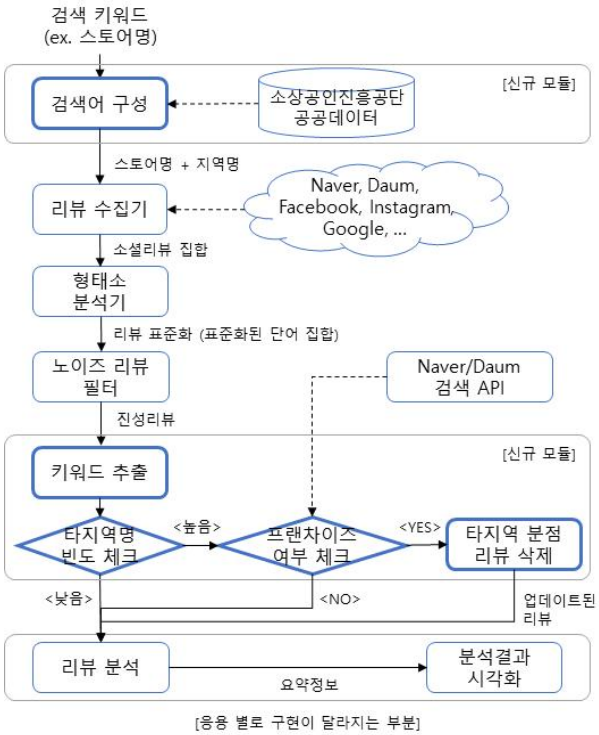

Fig. 3. Proposed method for social big data analysis reflecting characteristics of franchise stores

한 가지 주의할 점은 위 방법에서 FP(False-Positive) 오류가 발생할 수 있다는 점이다. 이는 제안 방법에서 타 지역의 분점 관련 지역명이 리뷰에서 발견되면 노이즈로 간주하고 일괄적으로 삭제한다는 점에 기인한다. 이에 반해 삭제된 리뷰 중에는 타 지역명이 언급되긴 하나 실제 노이즈 리뷰가 아닌 경우도 있을 수 있다. 예를 들어, "구구양꼬치"의 경우, 제안한 방법에서는 울산이 아닌 "창원"이나 "부산", "수영" 등의 지명이 리뷰에서 언급되면 타 지역 분점을 언급하는 것으로 간주하고 삭제한다. 그러나 실제로는 울산 본점을 설명하는 과정에서 다른 분점의 존재가 잠깐 언급될 수 있다. "창원, 수영에도 새로 생겼다고 하는데 ...", "수영에서 다녀온 구구양꼬치, 알고 보니 울산이 본점 ..." 등의 리뷰들은 실제 노이즈가 아님에도 불구하고 타 분점의 지역명이 들어갔다는 이유만으로 단순 삭제된다.

다행히 실제 리뷰 데이터를 수집하여 실험한 결과, FP 오류의 비율은 노이즈로 판단된 리뷰 수 대비 평균 6.4% 정도로 조사되었다. 이는 전체 리뷰 수 대비 평균 0.36%에 해당하며, 전체 분석 정확도에는 큰 영향을 미치지 않음을 알 수 있다. 관련 내용은 아래에서 이어 설명한다.

IV. Performance Evaluation

제안 방법에 대한 정확도를 검증하기 위해, 온라인에서 수집된 실제 리뷰 데이터를 이용하여 실험을 수행하였다. 실험을 위해 울산의 각 지역구별로 20개씩, 총 100개의 프랜차이즈 스토어를 선정하였으며(Appendix A 참고), 해당 스토어들을 대상으로 온라인으로부터 수집된 3만여 건의 리뷰 중, 노이즈 필터링 단계를 거친 약 9,300여 건의 진성 리뷰를 실험에 이용하였다.

먼저 [그림 4]는 각 지역구별로 수집된 전체 리뷰수와 제안 방법에 의해 프랜차이즈 분점 관련한 노이즈 리뷰로 분류된 리뷰수와 비율을 보여준다. 전체 9,300여건의 리뷰 중 타 지역의 분점을 언급한 노이즈 리뷰 수는 516건이었으며, 전체의 5.6% 정도에 해당하였다. 이는 제안 방법을 이용할 경우, 노이즈 필터링의 정확도가 기존에 비해 약 5.6% 정도 향상될 수 있음을 의미한다.

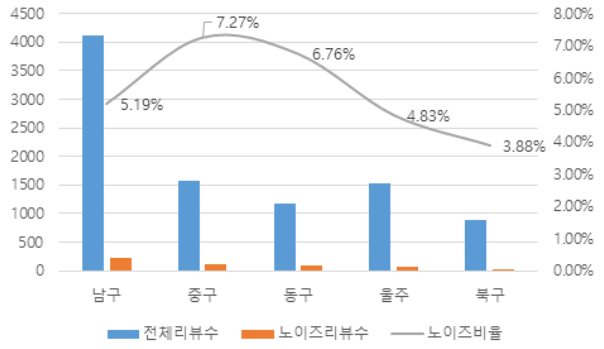

Fig. 4. Ratio of noise reviews detected from social reviews gathered from stores in Ulsan metro city

[그림 5]는 제안 방법에서 노이즈로 분류된 리뷰 중 FP(False-Positive) 오류로 판명된 리뷰 수와 비율을 보여준다. FP 오류 여부는 수작업으로 판별하였으며, 노이즈 리뷰 516건 중 33건이 FP 오류로 판명되었다. 이는 노이즈 리뷰수 대비 약 6.4%에 해당하며, 제안 방법의 노이즈 필터링 정확도가 평균 93.6% 정도임을 알 수 있다.

FP 오류 수를 전체 리뷰수와 대비하여 비교하였을 때, FP 오류 비율은 약 0.36%에 해당한다. 이는 소셜 리뷰 분석 대상이 되는 리뷰의 수가 9,300건임을 감안할 경우, 단 33건의 리뷰가 분석에서 제외되는 것이므로 실제 분석 결과의 정확도에는 거의 영향을 미치지 않는다고 볼 수 있다.

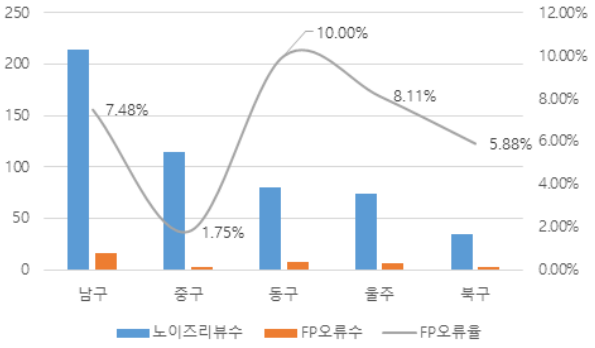

Fig. 5. Ratio of false-positive errors shown in noise reviews of Ulsan metro city data

다음으로 3.2절에서 설명한 검색어 설정 방법과 3.3절에서 설명한 프랜차이즈 분점 정보를 활용한 노이즈 필터링 방법에 대해, 각 방법을 통해 정확도가 어느 정도 향상될 수 있는지 비교하고자 하였다. 검색어 설정 방법만 적용했을 때 필터링되는 리뷰 수는 340건으로 전체의 3.66%에 해당하였으며, 이 중 FP 오류 수는 24건으로 조사되었다. 이에 반해, 분점 정보를 활용한 노이즈 리뷰 필터링 방법을 적용했을 때 필터링되는 리뷰수는 약 176건으로 전체의 1.9%에 해당했으며, 이 중 FP 오류 수는 9건이었다.

Table 2. Comparison of accuracy improvement for processing steps discussed in Section 3.2 and 3.3

	# of identified noise reviews	# of false-positive errors
Applying method of Section 3.2	340	24
Ratio of total reviews	3.66%	0.26%
Applying method of Section 3.3	176	9
Ratio of total reviews	1.9%	0.1%

위 실험 결과로부터 공공데이터 누락에 따른 노이즈 필터링 오류 비율이 전체 오류의 34% (=176/(340+176)) 정도로 예상보다 크게 나타났으며, 이는 프랜차이즈 스토어를 대상으로 소셜 빅데이터 분석에 있어 만족할 만한 정확도를 얻으려면 3.2절과 3.3절에서 제안한 방법을 모두 활용해야 함을 의미한다고 볼 수 있다.

V. Conclusion and Future Work

본 논문에서는 프랜차이즈 스토어를 대상으로 한 소셜 빅데이터 분석 방법에 대해 제안하였다. 제안 방법에서는 소상공인진흥공단에서 제공하는 공공데이터를 이용하여

검색어에 지역명을 어떻게 추가시킬 것인지와 관련한 알고리즘을 제안하였다. 그리고 공공데이터에서 누락된 정보로부터 발생할 수 있는 오류를 보완하기 위해, 네이버 및 카카오에서 제공하는 검색 API를 활용하여 프랜차이즈 분점 정보를 알아낸 후, 이를 기반으로 분석 대상이 아닌 타 지역 분점들의 리뷰를 걸러내기 위한 방법을 함께 소개하였다. 제안 방법을 검증하기 위해 온라인에서 수집된 실제 리뷰를 대상으로 실험을 수행하였으며, 실험 결과 노이즈 필터링 측면에서 제안 방법을 적용할 경우 기존에 비해 필터링 정확도가 평균 5.6% 향상됨을 알 수 있었다. 그리고 제안 방법의 노이즈 필터링 자체의 정확도는 평균 93.6%로 조사되었으며, 분석 결과의 정확도에는 미미한 영향을 주는 것으로 확인되었다.

제안 방법은 분석 대상이 아닌 타 지역 분점들의 리뷰를 걸러내기 위해 지역명을 기반으로 한 패턴 매칭을 수행한다. 패턴 매칭은 리뷰가 의미하는 문맥 내용을 충분히 반영하는데 한계가 있으므로, 진성 리뷰를 노이즈로 판단하고 걸러내는 False-Positive 오류를 발생시킬 수 있다. 따라서 분석 정확도를 더욱 높이기 위해서는 문맥 내용을 적절히 반영할 수 있는 기계학습 기반의 알고리즘 적용이 필요하다. 또한 다양한 지역의 프랜차이즈 스토어에 대해 제안 방법을 적용해 봄으로써, 제안 방법의 정확도를 더욱 높이기 위한 연구를 지속할 예정이다.

REFERENCES

[1] H. G. Kim, "Developing a Big Data Analysis Platform for Small and Medium-Sized Enterprises," Journal of the Korea Society of Computer and Information, Vol. 25, No. 8, pp. 65-72, Aug. 2020.

[2] W. L. Kang, H. G. Kim, and Y, J. Lee, "Reducing IO Cost in OLAP Query Processing with MapReduce," IEICE Trans. Inf. & Syst, Vol. E98-D, No. 2, pp. 444-447, Feb. 2015.

[3] Naver Open API, https://developers.naver.com/docs/common/openapiguide/

[4] Google Developer API, https://developers.google.com/

[5] H. G. Seo and H. W. Park, "Design and Implementation of Potential Advertisement Keyword Extraction System Using SNS," Journal of the Korea Convergence Society, Vol. 9, No. 7, pp. 14-24, 2018.

[6] O. J. Lee, S. B. Park, D. Chung, and E. S. You, "Movie Box-Office Analysis Using Social Big Data," Journal of the Korea Contents Society, Vol. 14, No. 10, pp. 527-538, 2014.

[7] C. Lee, D. Choi, S. Kim, and J. Kang, "Classification and Analysis of Emotion in Korean Microblog Texts," Journal of KIISE, Vol. 40, No. 3, pp. 159-167, Jun. 2013.

[8] J. Y. Chang, "A Sentiment Analysis Algorithm for Automatic Product Reviews Classification in Online Shop ping Mall," Vol. 14, No. 4, pp. 19-32, 2009.

[9] C. Park and C. Lee, “Korean Movie Review Sentimental Analysis using RNN-based Variational Inference,” Proceedings of the 2018 Korea Software Congress, pp. 587-589, December 2018.

[10] Y. Oh, M. Kim, and W. Kim, “Korean Movie Review Sentiment analysis Using Parallel Stacked Bidirectional LSTM Model,” Proceedings of the 2018 Korea Computer Congress, pp. 823-825, June 2018.

[11] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank,” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1631-1642, Seattle, Washington, USA, October 2013.

[12] K. S. Tai, R. Socher, and C. D. Manning, “Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks,” Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (IJCNLP), pp. 1556-1566, Beijing, China, July 2015.

[13] A. Mousa and B. Schuller, “Contextual Bidirectional Long Short-Term Memory Recurrent Neural Network Language Models: A Generative Approach to Sentiment Analysis,” Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 1023-1032, Valencia, Spain, April 2017.

[14] H. G. Kim, "Efficient Keyword Extraction from Social Big Data Based on Cohesion Scoring," Journal of the Korea Society of Computer and Information, Vol. 25, No. 10, pp. 87-94, Oct. 2020.

[15] Y. W. Yu and H. G. Kim, "Interactive Morphological Analysis to Improve Accuracy of Keyword Extraction Based on Cohesion Scoring," Journal of the Korea Society of Computer and Information, Vol. 25, No. 12, pp. 145-153, Dec. 2020.

[16] Z. Jin and K Tanaka-Ishii, "Unsupervised Segmentatino of Chinese Text by Use of Branching Entropy," The Journal of Korea Navigation Institute, pp. 428-435, Jul. 2006.

[17] H. J. Kim and S. J. Cho, "Cleansing Noisy Text Using Corpus Extraction and String Match," MS. Thesis, Seoul National University, 2013.

[18] E. Kim, "The Unsupervised Learning-based Language Modeling of Word Comprehension in Korean," Journal of the Korea Society of Computer and Information, Vol. 24, No. 11, pp. 41-49, Nov. 2019.

[19] M. Kim, S. Hong, and I. H. Suh, "Convolutional Neural Network Based Filtering-Scoring System for Rating Prediction of Travel Attractions using Social Media," Journal of the Institute of Electronics and Information Engineers, Vol. 56, No. 9, pp.

891-897, Sep. 2019.
[20] J. Yeon, J. Myung, J. Shim, and S. Lee, "Characteristic Set and Collaborative Filtering for Review Selection," Proceedings of the 2012 Korea Computer Congress, pp. 43-45, 2012.
[21] Kokoma, http://kkma.snu.ac.kr/documents/index.jsp
[22] Hannanum, http://semanticweb.kaist.ac.kr/hannanum/index.html
[23] JQCloud, https://mistic100.github.io/jQCloud/
[24] Small Business Promotion Agency, https://www.semas.or.kr

APPENDIX. A

Table 3. Franchise store names used for our experiments discussed in Section 4

울산 남구	울산 동구	울산 중구
808웨스트도어	골목식당 일산점	225토마토스트릿 호계점
고반식당 울산삼산점	김충기꽃삼겹	74족발 호계점
구구구양꼬치	녹색의향기 일산점	가리명가
대가야삼계탕 달동점	대한육회	감성대패 매곡점
돌부대찌개	라라코스트	교동면옥 울산점
디에이블 업스퀘어점	롤링파스타 일산점	대방낙지 명촌점
랑데자뷰 울산점	미미 일산점	도야족발 호계점
산청자매훠궈	바보형제쭈꾸미 일산점	돈오리 명촌점
생선구이생생	보교	라멘집입니다 호계점
스시아오 삼산점	봉창이칼국수	라화공방 송정점
스컹크웍스 삼산점	산너머남촌 주전점	명촌순두부보쌈
아카렌	언양닭칼국수	바다바라기
오별난멸치국수	유황먹은족발	사계진미 신천점
유동커피	윤회네깃대삼겹살	삼정숯불갈비 명촌점
율촌갈비탕	제갈콩명	스시아오 화봉점
정씨함박	제주은희네해장국	역전할머니맥주
진미간장게장	종로쭈꾸미	와우족발
포항황소곱창	코모레비	정안정
홈즈앤루팡	쿠우쿠우 울산동구점	정코다리 울산북구점
홉스피제리아 삼산점	해동중화요리	화로상회 울산명촌점

울산 울주군	울산 중구	
1984나폴리 울산점	225토마토스트릿 태화	
225토마토스트릿 범서	고인돌삼겹살 태화점	
갈비구락부	꼬꼬사우나찜닭 태화점	
덤덤덤쪽갈비 울산남창	녹색의향기 우정혁신점	
동부숯불왕갈비 언양점	더치앤빈	
서울칡냉면 언양점	도야족발 서동점	
소문난아구찜	등대갈비	
소문난왕뽈찜	라멘집입니다 울산성안	
언양닭칼국수 덕신점	방이편백 울산태화점	
역전할머니맥주 구영리	브라운곰돈까스	
오여진족구이 언양점	샤브나인 태화점	
울산매운수제비집	성서본가막창	
일류숯불고기 언양점	알통떡강정	
제주은희네해장국 범서	은화수식당 울산성남점	
진송추어탕 울산언양점	인생식당 성안점	
짬뽕상회 덕신점	조마루뼈다귀 태화점	
철판떼기	종로쭈꾸미 성안점	
킹부대찌개 언양점	코지누크	
트레비어 언양점	함양집	
한우리 언양점	호원갈비 유곡점	

Authors

Hyeon Gyu Kim received the B.S. and M.S. degrees in Computer Science from University of Ulsan, and Ph.D. degree in Computer Science from Korea Advanced Institute of Science and Technology, Korea, in 1997, 2000 and 2010, respectively. Dr. Kim joined the faculty of the Division of Computer Science and Engineering at Sahmyook University, Seoul, Korea, in 2012. He is currently an Associate Professor in the Division of Computer Science and Engineering, Sahmyook University. He is interested in big data processing, data stream processing, and mobile computing.