텍스트마이닝을 이용한 청소년의 성고민 분석

유정옥

동아대학교 간호학부 교수

Text Mining for Identifying Topics in Internet Q&A about Adolescents' Sexual Concerns

Jungok Yu

Professor, College of Nursing, Dong-A University, Busan, Korea

Objectives: Because of its accessibility and anonymity, the internet is common way for adolescents to find information about sensitive issues, such as sexual concern. The purpose of this study was to explore the concerns of adolescents about sex through internet Q&A on a Korean popular portals site by applying a text mining. **Methods:** The web page of the NAVER Q&A was crawled by searching the keyword using R. A total of 3,150 questions (male 1,440, female 1,710) from 2003 to 2017 on adolescent sexual concerns were used as object. Latent Dirichlet Allocation (LDA) was applied to perform text mining on the dataset. **Results:** In the LDA, 7 topics were selected for male adolescents and 11 topics for female adolescents such as same sex attraction, sexual identity confusion, masturbation, genital anatomy and sexual assault. **Conclusions:** This study provides useful information on sexual concerns of Korean adolescents. Topic modeling could be used as a tool to understand the online big data by discovering relevant topic.

Key words: Text mining, Topic modeling, Adolescent, Sexual health, Internet

서론

성관계 경험이 있다고 응답한 청소년은 전체의 5.7%로 나타났으며, 성관계 경험자 중 성관계 시작 평균 연령은 만 13.6세로 조사되었다[1]. 청소년기의 성경험은 10대 임신 및 그로 인한 인공임신중절수술을 경험, 성병 발생 위험을 증가[2]시켜 청소년의 생식건강을 위협한다. 실제 성경험이 있는 국내 청소년 14명 가운데 1명(7.3%)은 성매개 감염병에 걸린 경험이 있는 것으로 보고되고 있으며[2], 성경험이 있는 국내 여자 청소년 중 10.5%가 임신을 해 본 적이 있다고 하였고 이 중 81.0%는 임신중절을 경험한다[1]. 이런 부정적인 측면을 예방하고자 교육부에서는 학생들에게 성에 대한 올바른 지식과 정보를 제공하고 이를 통해

바람직한 성가치관과 성행동을 할 수 있도록 학교 내에서 체계적인 성교육을 실시하도록 하고 있으며 최근엔 국가수준의 성교육 표준안을 개발하여 제시한 바 있다[3].

그러나 청소년들의 성교육 이수 실태를 보면, 청소년의 71.9%가 최근 1년 동안 수업, 또는 방송교육 등을 통해 학교에서 성교육을 받은 적이 있다고 보고하였지만[1], 그 이면에는 주 1회 교육을 받은 경험은 2.7%, 월 1회 교육 경험은 8.7%로 그 외엔 학기에 1-2회 또는 1년에 1회 정도 성교육을 받은 것으로 나타나 실제적인 교육이 이루어지지 않고 있었다[4]. 수업 방법도 시청각, 또는 강사 특강이 대부분이어서[4] 초등학교에서 고등학교로 학교급이 올라갈수록 학교 성교육의 도움되는 정도가 유의미하게 낮아지는 경향을 보였는데[5] 이에 학생들은 ‘학

Corresponding author: Jungok Yu

32 Daesingongwon-ro, Seo-gu, Busan 49201, Korea
Tel: +82-51-240-2783, E-mail: joyu@dau.ac.kr

Received: April 09, 2019 Revised: May 15, 2019 Accepted: May 22, 2019

*This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1C1B5017614).

No potential conflict of interest relevant to this article was reported.

How to cite this article:

Yu J. Text mining for identifying topics in internet Q&A about adolescents' sexual concerns. J Health Info Stat 2019;44(2):181-188. Doi: <https://doi.org/10.21032/jhis.2019.44.2.181>

© It is identical to the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0>) which permit unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2019 Journal of Health Informatics and Statistics

생들이 알고 싶은 내용 중심', '충분하고 정확한 지식을 주는 수업' 등이 필요하다고 응답하였다[4]. Lee [5]의 연구에서도 청소년들이 가장 알고 싶다고 하는 주제는 상황별 대처 방법이라고 하였는데 구체적으로 청소년들이 성과 관련하여 어떤 문제에 직면하고 있는지 조사된 바가 거의 없어 청소년들이 궁금해 하는 성에 관한 이슈에 대해 조사하여 요구에 부합하는 실질적인 교육을 제공할 필요가 있다.

개개인이 성에 관해 가지는 질문은 다양한데 기존 설문지 등을 통한 조사방법은 대상자가 원하는 답이 문항에 없을 경우 적당히 응답을 선택하여 잘못된 해석을 내릴 수 있고, 또는 질문과 응답 목록이 피상적이어서 문제의 핵심을 파악하지 못할 수도 있으며 솔직한 응답이 아니라 사회적으로 우세한 가치에 부합하는 응답을 하는[6] 한계가 있다. 특히 청소년의 성과 관련된 고민은 사회적으로 수용되지 못하는 부분들이 많아 설문지를 통해 파악하는 것은 많은 사실이 은폐될 가능성이 높을 것으로 생각된다. 사람들은 이상 증상을 느끼거나 건강 문제에 직면할 때 건강정보를 찾게 되는데, 많은 경우 인터넷을 통해 정보를 찾는다[7]. 특히 우리나라 청소년의 경우 99.9%에서 인터넷을 이용하고 있으며[8], 인터넷은 익명으로 의사소통이 가능하기에 청소년들은 성고민과 같은 민감한 정보의 경우 면대면을 통해 정보를 얻기보다 인터넷을 활용하는 경우가 더 많다[9]. 이런 맥락에서 볼 때 인터넷에 올려진 청소년의 성고민 글을 분석하면, 청소년의 실질적인 성관련 고민을 제대로 파악할 수 있을 것으로 기대된다.

국내 최대 포털 사이트인 네이버(NAVER) [10]의 경우 지식iN (지식인, <https://kin.naver.com/index.nhn>)을 통해 사용자가 올린 질문이나 궁금한 내용, 고민에 대해 다른 사용자들이 자발적으로 답을 달면서 지식을 주고 받도록 하고 있으며 최근에는 변호사, 의사, 상담기관 등 전문가 상담을 도입하여 신뢰도를 높이고 있다[11]. 이에 네이버 지식인의 Q&A와 같은 질문 코너에는 사용자들이 건강을 포함한 다양한 질문을 하고 있으며 대개 이 질문들은 생활의 실질적인 정보 요구를 표현하는 경우가 많다. 이런 인터넷 상의 건강정보 요구에 대한 의미 있는 양상이나 변화를 파악하기 위한 연구들이 시도되어 네이버 지식인에 올려진 건강상담[12] 분석 연구가 시도되었다. 청소년의 사이버 성상담 사례에 대한 분석 연구는 국내에서는 Park et al. [13]의 연구에서 청소년 성상담 센터의 3년간 임신/피임 및 자위에 대해 상담 내용을 성별, 중고등학생별로 분류하여 분석을 시도한 연구가 있으나 특정 주제에 국한하여 분석하였다는 한계가 있다. 국외에서는 전화 또는 온라인 청소년 성상담 내용을 질적 내용 분석[14,15]하여 청소년의 성고민을 파악하였다. 그러나 현재까지 진행된 이 연구들은 수작업으로 몇 백 건의 내용을 파악[13-15]하였다는 한계와 일부는 병원 홈페이지 상담 게시판 자료를 분석하였기[15] 때문에 서비스 이용 비용, 서비스 이용 시 부모 동의 등 고민의 내용이 병원 이용과 관련된 내용으로 치우

쳐 있다는 제한점이 존재한다. 또한 국외 연구의 경우 사회문화적 차이로 인해 국내 청소년의 성고민과는 차이가 있을 것으로 생각되어 국내 청소년 대상 성고민의 주제를 파악할 필요가 있다.

최근 기존의 사회과학적 조사방법의 문제를 보완하기 위한 방법으로 데이터 마이닝(data mining)이 적용되고 있으며 그중 텍스트 마이닝(text mining)은 비구조화된 대량의 텍스트로부터 패턴을 찾아내어 텍스트로부터 이전에 알려지지 않은 숨은 지식을 제공한다[16]. 청소년의 성고민과 관련된 요구도 인터넷 상의 Q&A 분석을 통해 실질적인 요구 사항을 파악할 수 있을 것으로 생각되며 이를 통해 추출된 개념과 관련된 내용들은 이후 청소년을 위한 성건강 프로그램 개발에 기초자료가 될 것으로 기대된다.

본 연구의 목적은 텍스트 마이닝의 토픽클러스터링(topic clustering) 분석 방법을 적용하여 네이버 지식인에서 청소년의 성고민과 관련된 인터넷 Q&A 내용을 탐색하여 청소년의 성고민 주제를 도출하고자 하였다.

연구 방법

연구대상 및 자료

본 연구는 한국 청소년의 성고민과 관련된 내용을 추출하기 위해 국내 검색엔진 중 점유율 1위인 네이버 지식인의 질문 내용을 대상으로 하였다. 네이버 지식인 서비스가 2002년 10월 7일부터 시작은 되었으나 서비스 초기여서 관련 자료가 거의 없는 점을 감안하여 본 연구의 자료 수집은 2003년 1월 1일부터 2017년 12월 31일까지 네이버 지식인 지식 Q&A에 올려진 청소년의 성고민과 관련된 질문 제목과 질문 내용을 분석 대상으로 하였다.

네이버 지식인 Q&A에서 키워드가 포함된 질문 제목과 내용을 수집하기 위해 `httr`, `rvest`, `stringr`의 R 패키지를 이용하였다. 네이버 지식인 검색 화면에서 키워드로 '고등학생 or 중학생 or 여고생 or 여중생 or 남고생 or 남중생 or 십대 or 10대 or 청소년' and '성' and '고민'으로 하고 제외어로는 '편입, 진로, 직업, 알바, 탈모, 작명, 이름풀이'를 추가하여 검색하였다. 제외어를 선정하게 된 이유는 한국어는 소리글자로 키워드 검색어로 검색을 하였을 때 한국어의 동음이의어 특성에 따라 '성(性, sex)'과 '성(姓, family name)'이 명확히 구분되지 않기 때문에 '이름(name)'과 관련된 키워드가 들어간 문서가 함께 추출되어 제외하고자 하였다. 또한 청소년의 '고민'에 해당하는 편입, 진로 등도 많이 추출되어 검색 단계에서 제외어로 선정하였다. 2003년 1월 1일부터 2017년 12월 31일까지 키워드로 검색한 자료는 총 18,114건이었으며 이들 내용 중 중복 3,633건, 분석 내용이 아닌 자료 등 11,331건을 제외하여 최종 남자청소년 1,440건, 여자청소년 1,710건의 내용을 분석하였다.

자료 분석

텍스트 전처리

수집된 자료는 비구조화 문서이므로, 텍스트 전처리 과정을 통하여 정보 추출이 가능한 형태로 변환하는 작업이 필요하다. 특히 인터넷 Q&A에 올린 질문의 텍스트는 표준어로 표현되지 않았을 것으로 예상되어 비속어, 약어 등을 선별하고 정리할 필요가 있다. 데이터 전처리 작업을 위해서 먼저 형태소 분석(morphological analysis)을 통해 문서의 입력 문자열을 형태소열로 변환하는 작업을 수행하였다. 형태소 분석을 위해 한국어 자연어 처리(Natural Language Processing, NLP)를 위한 오픈소스 한글 형태소 분석기 은전한닢(<http://eunjeon.blogspot.kr>)을 R에서 사용할 수 있도록 도와주는 NLP4kcc패키지를 사용하였다. 형태소 분석 시 범용 사전에 의해 단어를 분할하게 되는데 이때 사용자의 의도와 다르게 분할되는 것을 방지하기 위해서는 사용자 사전을 구축하여 형태소 분석에 반영할 필요가 있다. 예를 들어 성정체성을 '성'과 '정체성'이라는 단어로 분리되지 않도록 하기 위해 사용자 사전을 만들어 '성정체성'이라는 단어를 등록하였다.

형태소 분석된 파일은 R 프로그램의 tm 패키지를 이용하여 데이터 전처리 작업을 실시하였다. 형태소 분석된 파일을 말뭉치(corpus)로 전환한 후 문장부호, 특수기호, 숫자를 제거하였고 분석에 불필요한 의미 없는 특정단어(예: 하다, 있다, 되다, 내공, 답변, 부탁 등)는 불용어 처리하여 삭제하였다. 동의어는 예를 들어 '톡', '카톡', '카카오톡' 또는 '야동', '음란물' 등은 같은 뜻을 가지므로 '톡', '카톡'을 '카카오톡'으로, '야동'을 '음란물'로 변경하는 등 동의어 처리를 하였다.

분석대상 단어 추출

데이터 전처리 단계를 거친 후 텍스트 마이닝을 실시하기 위해 여러 개의 형태소로 구성된 각각의 문서를 문서-용어 매트릭스의 형태(DocumentTermMatrix, DTM)로 저장하였다. DTM은 행에는 문서, 열에는 용어가 제시되는데, 용어 칸에는 분해된 단어가 제시되며 각 셀에는 문서에서 나타난 단어의 총 빈도수가 제시된다[19]. DTM 생성 시 한 글자 단어, 예를 들면 '저', '때', '것' 등의 단어는 분석에서 의미를 제공하지 못하여 필요가 없으므로 두 글자 이상의 단어로만 DTM이 생성되도록 하였다. 한 글자 단어 중 의미가 있는 단어들은 삭제되지 않도록 동의어 처리 단계에서 다른 단어로 대체하였는데, 예를 들면 '몸'은 '신체'로, '냉'은 '냉분비물'로 동의어 처리하여 삭제되지 않도록 하였다. 이런 과정을 거쳐 1,440건의 남자청소년 자료에서는 총 6,282개 단어가 추출되었으며 1,710건의 여자청소년 자료에서는 7,410개 단어가 추출되었다. 이들 단어 중 출현빈도가 낮은 단어를 선별하기 위해 문서의 97.0% 이상에서 출현하지 않은 단어(sparse term)를 삭제하여 최종적으로 남자청소년에서는 259개 단어, 여자청소년에서는 311개

단어가 추출되었다.

추출된 단어로 토픽 모델링을 수행하기 위해 단순히 빈도뿐만 단어의 중요도를 분석하는 것의 단점을 보완하고자 정보추출 분야에서 널리 알려진 단어중치기법인, 단어 빈도-역문서 빈도(Term Frequency-Inverse Document Frequency, TF-IDF)를 적용하였다. 이를 통해 특정 단어가 문서에서 갖는 중요도를 파악할 수 있다[17]. 즉 TF는 어떤 단어가 전체 문서 내에서 차지하는 출현 비중을 말하는데 일반적으로 TF가 클수록 문서에서 중요한 단어라고 볼 수 있다. 그러나 이 단어가 여러 문서에서 지나치게 빈번하게 사용되는 것은 그 키워드가 흔해 의미가 없다는 것을 뜻하기도 한다. 그러므로 TF뿐만 아니라 DF의 역수인 IDF도 단어의 가중치를 계산하기 위해 이용한다. IDF는 단어가 문서 집합 내에서 공통적으로 출현하는 비중을 표현하는 지표로, 빈번하게 나온 문서의 단어들을 역수로 표현하여 계산한 값으로 높은 IDF를 갖기 위해서는 특정 단어가 특정한 문서에서만 집중적으로 등장해야 한다. 어떤 단어가 사용되었을 때 어떤 주제를 가진 문서인지를 알기 위해서는 높은 TF 값을 갖는 것은 물론 동시에 높은 IDF 값을 가져야 한다[18]. TF-IDF는 TF와 IDF 값을 곱하여 산출한 값[17]으로 이를 적용하여 DTM을 생성하게 되면 문서들에서 매번 등장하는 흔한 단어들을 걸러낼 수 있게 된다. 본 연구에서는 토픽모델링 수행 전 추출된 단어에 TF-IDF를 반영한 DTM을 생성하여 TF-IDF 값이 하위 25% (여자 0.08 이하, 남자 0.1 이하)에 해당하는 단어를 삭제하여 DTM의 차원을 축소하였다. 이를 통해 토픽 모델링에 반영된 최종 분석 단어는 남자청소년 235개 단어, 여자청소년 239개 단어였다.

잠재 디리클레 할당(Latent Dirichlet Allocation [LDA]) 기반 토픽 분석

토픽 모델링은 최종 선정된 의미 정보를 기반으로 비구조적으로 모여 있는 문서들을 구조적으로 조직화하는 과정으로 문서 집합 내용의 유사도에 따라 여러 개의 소집단으로 분할하여 전체 데이터가 포함하는 주제나 성격을 파악할 수 있게 한다[16]. 이는 문서 집합 내 동시 출현 빈도가 높은 단어들을 기준으로 토픽을 생성하면 각 토픽 주제는 연구자 및 유관 전문가의 의견을 반영하여 명명이 된다. 본 연구에서는 토픽모델링 중 가장 많이 활용되고 있는 기법인 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) [18]으로 토픽분석을 실시하였다.

LDA로 토픽을 추출해내기 위해서는 토픽수를 사전에 결정해야 한다. 이 토픽수 결정은 연구자의 추정에 근거하여 투입되기에 해당 토픽수가 적정한지에 대해서는 정당화 논리가 병행되어야 한다[18]. 이에 본 연구에서는 적절한 토픽수 여부를 확인하기 위해 Intertopic Distance Map (IDM)을 생성하여 시각적으로 토픽 간 유사도를 확인하였으며 남자 7개의 토픽, 여자 11개의 토픽이 적절한 것으로 판단되었다. LDA로 도출할 수 있는 산출물은 토픽별 핵심 단어, 문서별 토픽 확률

Table 1. The topics of latent dirichlet allocation among boys and the most 15 probable words in the topics

	Label	Words	Ratio (%)
T1	Same sex attraction	friend, school, close, study, story, attend, time, hard, hang out, joke, problem, confession, life, middle school, wrong	17.8
T2	Gynecomastia	big, woman, breast, surgery, young, over-hang, body, put on, father, parent, mom, picture, exercise, worry, middle school student	12.4
T3	Masturbation	masturbation, premature ejaculation, ejaculation, come out, behavior, semen, porno, a day, coitus, method, big, time, treatment, ejaculate, high school student	14.2
T4	Sexual relation	girlfriend, coitus, woman, go out with, meet, kiss, high school student, ask, love, story, breast, boyfriend, sorry, touch, confession	12.9
T5	Genitals	genitals, erection, small, phimosis, circumcision, big, natural, size, glans penis, penis, sick, problem, condition, length, part	12.2
T6	Sexual identity	woman, gay, homosexuality, feel, boyfriend, sexual identity, friend, excitement, interest, love, confusion, cute, puberty, emotion, go mad	20.6
T7	Sexual impulse	older sister, sleep, erotic, sexual impulse, report, go mad, school, time, touch, sexual assault, head, attend, joke, help, vacation	10.0

값, 문서별 토픽번호 등이다. 이에 각 토픽을 나타내는 주요 핵심단어는 15개씩 추출하였으며 문서별 토픽 확률값과 토픽번호를 출력하였다. 해당 토픽 확률값이 높게 배당된 문서들을 확인하고 토픽별 추출된 단어의 연관성을 반영하여 각 토픽의 이름을 부여하였다.

연구 결과

청소년 성고민 토픽 분석

남자청소년 성고민의 토픽분석 결과는 Table 1과 같다. 분석 방법에서 제시한 바와 같이 토픽모델링에서 토픽수를 결정하는 것은 중요한 이슈로, 본 연구에서는 적절한 토픽수 결정을 위하여 토픽간 유사도를 IDM으로 시각화하여 확인하였다(Figure 1A). 남자청소년의 경우 7개의 토픽이 도출되었다. 각 토픽에서 도출된 단어 리스트와 문서에 매칭된 토픽 확률값을 활용하여 토픽별 주제를 명명하였다. 1번 토픽은 동성친구에게 성적 호감을 느낌으로 인한 문제, 2번 토픽은 여자처럼 유방이 커지는 문제, 3번 토픽은 자위와 조기사정의 문제, 4번 토픽은 성관계의 시기, 방법, 임신가능성에 관한 문제, 5번 토픽은 성기 크기, 포경 수술 여부에 관한 문제, 6번 토픽은 성정체성 혼란에 관한 고민, 7번 토픽은 성욕구 해소에 관한 고민으로 명명하였다. Figure 1A의 남자청소년의 IDM을 보면 토픽 간 분류가 잘 된 것으로 확인되며 1번 토픽과 4번 토픽에서 일부 중복되는 것으로 나타났다. 그러나 이것은 4번 토픽이 이성과의 관계에 관한 글이기에 만나고 고백하고 사귀는 일련의 과정이 1번 토픽의 동성 친구와의 관계와 관련된 부분과 중복되어 사용되는 단어가 있어서 그런 것으로 보이며 내용을 확인해 보았을 때 분명히 다른 주제로 확인되었다. 7개의 토픽 중 가장 많은 비율을 차지하는 고민 주제는 6번 토픽인 성정체성 혼란에 관한 주제로 20.6%를 차지하였고 두 번째는 1번 토픽으로 동성 친구를 좋아함으로 생기는 고

민이 17.8%, 세 번째는 3번 토픽으로 자위와 조기 사정에 관한 고민이 14.2%를 차지하였다.

여자청소년 성고민의 토픽분석 결과는 Table 2와 같으며 토픽 간 유사도 확인 결과는 Figure 1B와 같다. 여자청소년의 경우 11개의 토픽이 도출되었다. 각 토픽에서 도출된 단어 리스트와 문서에 매칭된 토픽 확률값을 활용하여 토픽별 주제를 명명하였다. 1번 토픽은 자위로 인한 우려, 2번 토픽은 성폭행당한 후 대처방법, 3번 토픽은 성정체성고민, 4번 토픽은 소음순의 크기, 색깔에 관한 고민, 5번 토픽은 생리의 불규칙성 또는 안 하는 이유, 6번 토픽은 동성친구에게 성적 호감을 느낌으로 인한 문제, 7번 토픽은 성관계에 대한 호기심, 성관계 시 피임법 등 고려점, 8번 토픽은 가족 내 성폭력 문제, 9번 토픽은 선생님을 짝사랑하는 고민, 10번 토픽은 연애 시 신체접촉을 포함한 진도의 문제, 11번은 가슴 사이즈에 관한 고민으로 명명하였다. Figure 1B의 여자청소년의 IDM을 보면 토픽 간 분류가 비교적 잘 된 것으로 확인되며 3번 토픽, 6번 토픽, 9번 토픽은 각각 성정체성고민, 동성친구에게 성적호감을 느껴서 발생하는 고민, 선생님을 짝사랑하는 고민으로 선생님을 짝사랑하는 고민에는 동성의 선생님을 좋아하는 감정에 대한 혼란이 포함되어 있어 일부 사용 단어가 중복되어 일부 겹치는 것으로 내용상으로는 분명히 구분되는 것으로 확인되었다. 또한 8번 토픽과 10번 토픽은 일부 겹친 부분이 있지만 8번은 가족 내 성폭력과 관련된 주제이고 10번은 남자친구와 신체접촉 진도를 어느 정도까지 해야 하는지, 거절을 어떻게 할지에 대한 고민이어서 내용상 다른 주제로 확인되었다. 11개의 토픽 중 가장 많은 비율을 차지하는 고민 주제는 3번 토픽의 성정체성고민으로 13.3%를 차지하였고 두 번째는 6번 토픽으로 동성에게 성적 호감을 느낌으로 인한 고민이 13.0%를 차지하였다.

Figure 1. Intertopic distance map.

Table 2. The topics of latent dirichlet allocation among girls and the most 15 probable words in the topics

	Label	Words	Ratio (%)
T1	Masturbation	masturbation, middle school student, older sister, method, picture, insert, finger, porno, in former days, search, got into, puberty, feeling, problem, ashamed	5.8
T2	Sexual assault	older brother, sexual assault, report, touch, parent, suffer, genitals, body, memory, phone call, contact, insert, private institute, dirty, get on	9.2
T3	Sexual identity	homosexuality, bisexuality, confusion, same sex, other sex, gay, heterosexuality, appear, romantic relationship, coming out, identity, surrounding, feeling, pretty, interest	13.3
T4	Labia minora	labia minora, surgery, stretch, severe, part, put on, vaginitis, parent, inconvenient, be born, say, hospital, worry, pants, skirt	7.2
T5	Menstruation	period, pregnancy, sick, worry, coitus, begin, leukorrhea, possible, anxiety, OBGY, hospital, symptom, ask, panty, late	7.0
T6	Same sex attraction	lesbian, friend, confession, interest, pretty, cute, know, jealousy, personality, entertainer, good feeling, tremble, hold hands, touch, palpitate	13.0
T7	Sexual relation	coitus, erotic, sexual impulse, excitement, experience, porno, feeling, marriage, pervert, high school student, curiosity, last, body, adolescent, regret	6.7
T8	Incest	mom, scary, sleep, father, die, younger brother(sister), help, occur, family, get out, sick, irritation, afraid, condition, ask	8.1
T9	Love for teacher	teacher, high school, study, student, personality, pretty, start, serious, class, different, behavior, care, help, appearance, reason	11.5
T10	Physical intimacy	boy friend, kiss, break up, text message, physical contact, touch, send, kakao talk, ask, phone, call, romantic relationship, fast, refusal	11.6
T11	Breast size	breast, big, small, put on, increase, way, exercise, touch, middle school student, body, worry, inform, serious, underwear, elementary school student	6.5

고 찰

본 연구는 청소년이 당면하고 있는 성고민을 파악하기 위해 청소년이 쉽게 접근하여 사용하고 있는 인터넷 포털 지식공유서비스 일종인 네이버 지식인에서 추출된 질문들로 텍스트 마이닝의 토픽분석을 시도하였다. 연구 결과, 청소년들의 다양한 성고민을 파악할 수 있었으며 일부는 학교 성교육에서 다루고 있지 못하는 부분들이 포함되었다.

이 중 남녀 청소년 모두에서 동성에 호감, 성정체성 혼란에 대한 부분이 같이 도출되었는데 이 부분은 최근 교육부에서 발표한 성교육 표준안[3]에서 사회적 합의가 이루어지지 않았다는 이유로 학교 내 성교육 주제로 포함시키지 못하도록 하고 있는 부분이다. ‘동성에 호감’과 ‘성정체성 혼란’에 관한 주제는 유사함에도 서로 주제가 분리된 이유는, 동성에 호감으로 명명한 토픽은 이성 친구를 좋아하는 것처럼 정서적, 성적으로 동성에게 끌리는 그 자체에 대한 고민으로 함께 있고 싶고 친해지고 싶는데 어떻게 친해질 수 있을지에 대한 고민과 좋아한다고 고백하고 싶지만 그런 행동이 동성 간에서는 사회문화적으로 이해받지 못하는 행동으로 생각되어 차마 좋아한다고 말하면 오히려 관계가 악화될 것에 대한 우려, 그래서 좋아하는 마음을 감추고 일상적인 동성 친구 사이처럼 지내지만 진심을 드러내지 못함에 대한 관계에 대한 고민 내용이 주를 이룬다. 반면, 성정체성 혼란으로 명명한 토픽은 남들과 달리 동성에게 매력을 느끼는 본인의 모습에 대한 의식과 지각으로 ‘지금 고민은 내가 동성애자일까 아닐까이다. 친한 여자가 달리 보이는데 사실 잘 모르겠다’, ‘남학생인데 남자친구가 갑자기 연애 대상으로 보인다. 게이인가, 정상인가?’, ‘남성에게 매력을 느낀다는 게 웬지 자괴감이 드네요’와 같이 동성에 대한 특별한 감정에 대해 혼란스럽고 당황해하며 사회적으로 비난받고 부정적으로 인식되는 동성에 성향을 지니는 것이 아닌가 하여 불안해 하는 고민글이 다수를 이룬다. 전국의 중고등학교 청소년 2,280명을 대상으로 청소년 동성애자 현황을 조사한 연구[19]에 의하면 동성에 성향이 있지 않을까 고민해 본 청소년은 응답자의 11.0%로 청소년의 성정체성 문제는 극히 일부의 청소년이 경험하는 것이 아니라, 청소년의 성 발달의 주요한 부분으로 다루어져야 함을 의미한다. 그러나 교육부의 성교육 표준안[3]은 성교육에 ‘다양한 성적 지향과 성 정체성에 대한 이해’에 관한 내용이 포함되어야 한다는 유엔(UN)과 휴먼라이츠워치의 권고[20]를 오히려 어기고 있다. 더불어 우리나라는 동성애에 대한 사회적 수용도가 다른 나라와 비교해 보았을 때 상당히 낮고[21] 이를 반영하여 학교 환경 또한 전반적으로 동성애에 대해 적대적인 특성을 보여 청소년들은 성정체성에 대한 고민을 숨기고 제대로 도움을 받지 못하고 있는 것으로 나타났다[22]. 그러므로 청소년의 ‘다양한 성적 지향과 성 정체성에 대한 이해’에 대한 교육은 전체 청소년을 대상으로 성정체성 형성이 성

적 발달의 자연스런 과정이라는 것을 알려줌으로써, 청소년들이 이성애든 동성애든 자신의 성적 지향에 대해 점검하고 타인의 성적 지향에 대해서도 존중할 수 있는 태도를 함양하는 교육이 필요하다[22].

남녀 청소년에서 공통적으로 나오는 다른 성고민 주제는 사춘기의 신체 변화와 생식기의 해부학적 형태에 관한 것이다. 남자청소년의 경우 생식기 부분에서는 음경 크기가 작은 것 같은데 정상인지, 포경 수술을 해야 하는지 여부가 포함되었고 여성형 유방인 것 같은데 수술을 해야 하는지에 관한 주제들이 나타났다. 여자청소년의 경우 생식기 부분에서 소음순의 색, 늘어남, 비대칭 등에 대한 생식기 관련 주제가 하나 도출되었고 가슴(유방) 크기가 작거나 큰데 어떻게 하면 크기를 줄이거나 크게 할 수 있는지에 대한 주제가 하나 도출되었다. 사춘기의 이차 성징과 생식기의 해부학적 형태에 대해서는 학교 보건교육 및 성교육에서 많이 다루고 있는 주제이나 본 연구에서 특징적으로 발견한 주제는 남학생의 여성형 유방과 여학생의 소음순에 대한 관심이다. 건강보험심사평가원 자료[23]에 따르면 가슴이 지나치게 큰 유방비대증으로 진료 받는 남성은 꾸준히 증가하는 것으로 나타났으며 남성 환자의 45.9%가 10-20대에서 나타나는 것으로 집계되었다. 남성의 유방비대증은 청소년기의 경우 성호르몬의 일시적인 불균형으로 발생하며 이런 경우 대개 성인이 되기 전에 사라지는 것으로 알려져 있어[24] 청소년 시기의 치료는 권장되지 않는다. 다만, 외모로 인한 자존감 저하, 우울 등의 심리적 문제[25]가 있을 수 있는 만큼 사춘기의 신체발달과 이차성징에 대한 교육을 할 때 남자의 성장 발달 과정에서 드물지 않게 일어나는 변화라는 사실을 미리 알려줄 필요가 있겠다. 여학생의 소음순에 대한 고민은 청소년이 가지고 있는 잘못된 성지식으로 인해 생기는 고민으로 왜곡된 생식기 이미지를 기준으로 자신의 소음순 모양, 색이 비정상이라고 고민하는 것으로 판단된다. 사춘기가 지나면 소음순이 늘어나며 그 모양 또한 제각각이고 비대칭도 정상이며 사춘기 동안 멜라닌의 변화로 색 또한 짙어지게 되는 것이 정상 발달[26]이라는 정보 또한 사춘기의 신체발달과 이차성징에 대한 교육을 할 때 포함시킬 필요가 있겠다.

세부 내용은 다르지만 남녀 청소년 공통적으로 도출된 또 다른 주제들은 자위, 성관계 등 청소년의 성행동과 관련된 주제들이다. 청소년의 자위는 이 시기에 빈번히 관찰할 수 있는 성행위의 한 형태로 국내 청소년의 자위행위 경험률은 남자청소년 88.1%, 여자청소년 59.7%로 보고[27]된 바 있다. 청소년 시기의 자위행위는 위험부담 없이 안전하게 자신의 성적 자아를 발견할 기회를 제공해 주지만, 알려진 바가 많지 않아 자위행위에 대한 잘못된 인식, 잘못된 정보는 청소년 시기에 흔히 나타나는 문제이다. 본 연구에서 남녀 청소년들 모두 실제적인 성관계에 대한 고민글(남 T4, 여 T7) 또한 한 주제로 분류되었는데 남녀 청소년별로 세부 내용은 조금 차이가 있었다. 남학생의 경우 성관계를

하고 싶다는 마음과 절제해야 한다는 마음 사이의 갈등, 상대와 신체 접촉의 진도가 허용되는 범위인지, 성관계를 했는데 임신 가능성이 있는지 등과 관련된 내용이었다. 여학생의 경우도 음란물, 친구를 통한 정보를 통해 성에 대한 관심과 성욕을 느끼게 되면서 그런 자신의 모습이 비정상인지 하는 것과 어느 정도의 나이면 성관계를 가져도 되는지에 대한 질문, 성관계 후 임신 가능성에 대한 질문 정도로 나누어졌다. 여학생의 경우 사귀는 사람과의 신체 접촉의 진도를 어느 정도까지 해야 할지, 어떻게 거절 의사를 표현할지에 대한 고민은 별도 토픽(여 T10)으로 분리되었다. 국가 성교육 표준안[3]에서는 청소년 시기의 성욕구를 인정하고 성욕구 및 성욕구 해소라는 교육 주제를 다루고는 있으나 성욕구 해소 방법이 건전한 취미를 가진다든가 운동을 한다든가 하는 금욕과 절제 위주의 내용으로 자위에 관한 부분은 용어 자체도 언급되지 않은 상황이었어서 학교 성교육 내에서는 청소년들이 실질적으로 당면하고 있는 문제에 대한 해답을 구하기 어려워 보인다. 이런 접근 방법은 성교육 표준안과 학교 성교육의 관점이 청소년의 성을 자연스러운 권리의 차원으로 이해하기보다 청소년의 성을 무성으로 간주하고, 성교육의 취지를 청소년의 성적 일탈행동으로 인한 범죄 예방 차원으로 보고 있기 때문으로 성교육을 바라보는 관점의 전면적인 검토가 필요하다고 생각된다.

마지막으로 남녀 청소년 모두에서 강제성추행, 성폭력과 관련된 주제가 도출되었다. 남학생(T7)의 경우 성추행을 강하게 느끼는 상황에서 신체 접촉을 했는데 이후 어떻게 대처해야 하는지에 대한 고민, 반면 성추행 또는 성폭행을 당했는데 어떻게 대처하고 신고해야 하는지에 대한 질문이 함께 있었다. 여학생의 경우 성폭행과 관련된 주제가 두 개로 분리(T2, T8)되어 도출되었는데 하나는 교제 중이거나 알고 있는 사람에게 강제로 성관계를 당했는데 어떻게 해야 하는지에 대한 주제와 가족 내에서 벌어지는 근친상간 문제에 대한 고민이었다. 교육부는 초·중·고 학년별로 연간 15시간 이상 의무적으로 성교육을 하되 그중 성폭력 예방교육 3시간을 포함하여 실시하도록 하고 있다. 성폭력이 무엇인지 사례, 사례별 대처법, 예방법 위주의 교육이라 발생 이후의 대처, 신고에 대한 문제 해결에 대해 미숙하여 관련 질문들이 많이 올라온 것으로 보인다. 성폭력 예방교육에 발생 이후에 대처 방법에 관한 교육도 필요하다고 생각된다.

본 연구는 청소년 전반적인 성고민에 대하여 연구자의 주제 범주에 대한 가정 없이 내용을 추출하고자 성, 고민이라는 키워드를 이용하여 자료를 추출하였다. 추출된 내용들은 청소년의 성고민에 대한 전반적인 내용이 다루어진 것으로 보이나 청소년이 올린 글에서 성, 고민이라는 단어가 포함되어 있지 않은 내용일 경우 추출 자체에서 제외되어 분석에서 빠졌을 가능성이 있다. 또한 익명성이 보장되는 인터넷 글의 특성상 연령이나 성별 등의 개인 정보를 알 수 없었다. 다만 질문의 패

턴상 도입 부분에 본인을 소개하면서 연령대나 성별이 드러나는데 청소년의 성고민에 한정하여 분석하기 위해 질문 글 내에서 질문자의 연령이나 성별 정보를 알 수 없을 경우 모두 분석에서 제외하였다. 그러한 과정에서 실제 청소년이 올린 다수의 글들도 분석에서 제외되었을 가능성이 있어 연구 결과를 해석하는 데 주의를 요한다. 또한 많은 청소년들이 인터넷을 이용하지만 그렇다고 게시판에 상담글을 올린 청소년이 대다수의 청소년의 성고민을 대표하지는 못한다. 그럼에도 다수가 아닐지라도 구조화된 설문지나 익명성이 보장되지 않는 상담에서 다룰 수 없는 청소년의 성고민에 대한 유용한 정보를 제공하였다는 데 본 연구의 의의가 있다.

결론

본 연구는 청소년이 실질적으로 당면하고 있는 성고민을 파악하기 위해 인터넷에 올려진 비정형 자료를 텍스트 마이닝을 이용하여 분석을 시도하였고 다량의 텍스트 자료에서 청소년의 성고민에 대한 주제를 파악하여 청소년의 성 건강에 관심을 가지는 연구자, 교육자, 정책 입안자에게 유용한 정보를 제공하였다는 점에서 의의가 있다. 다만 인터넷에 게시된 자료가 모든 청소년을 대표한다고 할 수는 없으므로 텍스트 마이닝을 통해 탐색된 정보를 기반으로 청소년 대상 심층면담 또는 설문지를 이용한 조사 연구를 수행하기를 제안한다.

ORCID

Jungok Yu <https://orcid.org/0000-0001-8062-0207>

REFERENCES

1. Ministry of Education, Ministry of Health and Welfare, Korea Centers for Disease Control and Prevention. The 12th Korea Youth Risk Behavior Web-based Survey. Sejong: Ministry of Education, Ministry of Health and Welfare, Korea Centers for Disease Control and Prevention; 2016 (Korean).
2. Lee SY, Lee HJ, Kim TK, Lee SG, Park EC. Sexually transmitted infections and first sexual intercourse age in adolescents: The Nationwide Retrospective Cross-Sectional Study. J Sex Med 2015;12(12):2313-2323. Doi: 10.1111/jsm.13071
3. Ministry of Education. Sex education guideline. Available at <http://www.schoolhealth.kr/web/bbs/selectBBSList.do?bbsTyCode=BBST01&bbsId=0108&lstnum1=2641&pageIndex=1&pageUnit=10&dataType=&search>

- chCnd=&searchWrd= [accessed on May 28, 2018].
4. Kim EJ, Ha SM. A study on sexual consciousness, the actual condition of school sex education and its need in high school students. *J Child Welf Dev* 2016;14(2):19-39 (Korean).
5. Lee GY, Song SH. Influencing factors on sexual knowledge among elementary, middle, and high school students. *J Korean Acad Nurs* 2015; 21(3):406-416 (Korean). Doi: 10.5977/jkasne.2015.21.3.406
6. Lee EO, Im NY, Park HA, Lee IS, Kim JI, Bae JE, et al. *Nursing research and statistics*. 1st ed. Paju: Soomoonsa; 2009, p. 241 (Korean).
7. Escoffery C, Miner KR, Adame DD, Butler S, McCormick L, Mendell E. Internet use for health information among college students. *J Am Coll Health* 2005;53(4):183-188.
8. Ministry of Science and ICT, Korea Internet and Security Agency. 2017 Survey on internet usage. Gwacheon: Ministry of Science and ICT, Seoul: Korea Internet & Security Agency; 2018 (Korean).
9. Valkenburg PM, Peter J. Online communication among adolescents: an integrated model of its attraction, opportunities, and risks. *J Adolesc Health* 2011;48(2):121-127. Doi: 10.1016/j.jadohealth.2010.08.020
10. BizSpring Inc. Search engine influx characteristics. Available at <http://www.internettrend.co.kr/trendForward.tsp> [accessed April 28, 2018].
11. NAVER Jisik-iN. Available at <https://ko.wikipedia.org/wiki/%EC%A7%80%EC%8B%9DiN> [accessed on May 28, 2018].
12. Kim YK, Lim BM. Internet health counseling for Korean medicine in the NAVER Jisik-iN. *Korean J Orient Med Prev* 2013;17(1):51-63 (Korean).
13. Park YS, Cho AM, Kim HW, Song SH, Kang YB, Cho YJ, et al. An analysis on cyber sexuality counseling case of youth. *J Future Orient Youth Soc* 2005;2(1):17-27 (Korean).
14. Kang M, Quine S. Young people's concerns about sex: unsolicited questions to a teenage radio talkback programme over three years. *Sex Educ* 2007;7(4):407-420. Doi: 10.1080/14681810701636010
15. Buzi RS, Smith PB, Barrera C. Talk with Tiff: teen's inquiries to a sexual health website. *J Sex Marital Ther* 2015;41(2):126-133. Doi: 10.1080/0092623X.2013.857375
16. Back YM. *Text-mining using R*. 1st ed. Seoul: Hanulmplus; 2017 (Korean).
17. Silge J, Robinson D. *Text mining with R: a tidy approach*. 1st ed. Sebastopol, CA: O'Reilly Media; 2017.
18. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003;3:993-1022.
19. Ham IH, Jee SH, Kim ST, Jung YH, Lee YS. *The worries of youth: Am I a homosexual?*. Busan: Korea Youth Counseling and Welfare Institute; 2003 (Korean).
20. Letter to the government of South Korea on the need to recognize sexual orientation and gender identity in sex education. Available at <https://www.hrw.org/news/2015/05/03/letter-government-south-korea-need-recognize-sexual-orientation-and-gender-identity> [accessed on May 28, 2018].
21. Global acceptance of homosexuality. Available at <http://www.pewglobal.org/2013/06/04/global-acceptance-of-homosexuality/> [accessed on May 28, 2018].
22. Kang BC, Kim JH. *A study of the lives of sexual minority youths*. Sejong: National Youth Policy Institute; 2006 (Korean).
23. HIRA Web Magazine. I'm a man, but I keep getting breasts. Is it female breast?. Available at <http://hirawebzine.or.kr/11236> [accessed on May 28, 2018].
24. Lemaire V, Cayci C, Simmons PS, Petty P. Gynecomastia in adolescent males. *Semin Plast Surg* 2013;27(1):56-61. Doi: 10.1055/s-0033-1347166.
25. Kinsella JC, Landfair A, Rottgers SA, Cray JJ, Weidman C, Deleyiannis FWB, et al. The psychological burden of idiopathic adolescent gynecomastia. *Plast Reconstr Surg* 2012;129(1):1-7. Doi: 10.1097/PRS.0b013-e3182361efd
26. Scarleteen. Why are my labia so weird?. Available at http://www.scarleteen.com/article/advice/why_are_my_labia_so_weird [accessed on May 28, 2018].
27. Kim HS. Masturbation of Korean adolescents. *J Korean Acad Psychiatr Ment Health Nurs* 2001;10(2):137-147 (Korean).