

교육정책관련 여론탐색을 위한 소셜미디어 감정분석 연구*

정진명**, 유기영***, 구찬동****

요약

웹2.0시대와 더불어 소셜미디어 서비스의 발달로 전통적인 여론형성의 기능이 매스미디어에서 소셜미디어로 일부 이동되었으며, 이런 현상은 계속 확대되고 있어, 정부 정책에 대한 소셜미디어 기반의 여론이 관심을 받고 있다. 특히, 교육정책은 다양한 이해관계자들이 존재하고, 정책의제 설정과정에서도 많은 의견의 충돌이 발생되기 때문에 정책을 수립함에 있어 대중의 여론을 파악하는 것이 더욱 중요하다고 할 수 있다. 본 연구는 교육정책관련 소셜미디어 서비스를 통해 작성된 문서들을 오피니언 마이닝 기법으로 분석하여, 교육관련 정책에 대한 대중의 여론탐색을 목적으로 하였다. 이를 위하여 소셜미디어 서비스를 통해 사용자들이 생산하는 교육정책 관련 문서들을 키워드 기반으로 수집하고, 토큰화 시킨 후 감성자질을 추출하고 감성사전으로부터 해당 문서의 감정을 점수화 하여, 특정 교육정책 키워드에 대한 대중의 여론을 탐색하였다. 그 결과 디지털교과서, 이러닝 등을 키워드로 하는 스마트교육 정책에 대해서는 긍정보다 부정적인 감정이 많은 것으로 나타났으며, 코딩교육, 컴퓨터적 사고 등을 키워드로 하는 소프트웨어 교육 정책에 대해서는 긍정적인 여론의 방향으로 나타났다. 자유학기제, 창의 인성교육 등을 키워드로 하는 일반 정책에 대해서는 부정적인 여론이 많은 것으로 나타났다. 또한 전체 분석 대상 문서 중에서 감정 자질이 전혀 추출되지 않은 문서가 20%나 되어 블로그나 트위터의 내용에 저자의 의견이 반영되지 않은 내용이 아직 일정 비율로 존재함을 알 수 있었다.

주제어: 교육정책, 감정분석, 소셜미디어, 여론탐색

A Study on Social Media Sentiment Analysis for Exploring Public Opinions Related to Education Policies

Chung, Jin-Myeong, Yoo, Ki-Young, Koo, Chan-Dong

Abstract

With the development of social media services in the era of Web 2.0, the public opinion formation site has been partially shifted from the traditional mass media to social media. This phenomenon is continuing to expand, and public opinions on government policies created and shared on social media are attracting more attention. It is particularly important to grasp public opinions in policy formulation because setting up educational policies involves a variety of stakeholders and conflicts. The purpose of this study is to explore public opinions about education-related policies through an empirical analysis of social media documents on education policies using opinion mining techniques. For this purpose, we collected the education policy-related documents by keyword, which were produced by users through the social media service, tokenized and extracted sentimental qualities of the documents, and scored the qualities using sentiment dictionaries to find out public preferences for specific education policies. As a result, a lot of negative public opinions were found regarding the smart education policies that use the keywords of digital textbooks and e-learning; while the software education policies using coding education and computer thinking as the keywords had more positive opinions. In addition, the general policies having the keywords of free school terms and creative personality education showed more negative public opinions. As much as 20% of the documents were unable to extract sentiments from, signifying that there are still a certain share of blog posts or tweets that do not reflect the writers' opinions.

Keywords: education policy, sentiment analysis, social media, exploring public opinion

2017년 11월 17일 접수, 2017년 12월 1일 심사, 2017년 12월 20일 게재확정

* 본 논문은 한국교육학술정보원에서 수행한 연구보고서(교육정책 아젠다 발굴을 위한 소셜미디어 빅데이터 분석 연구, 2016)의 일부를 수정·보완한 것이다

** 한국교육학술정보원 선임연구원(jinmong@gmail.com)

*** 교신저자, 경북대학교 IT대학 컴퓨터공학부 교수(yook@knu.ac.kr)

**** 한국교육학술정보원 연구원(cdkoo21@gmail.com)

I. 서론

소셜미디어는 SNS를 포함하는 더 큰 범주의 개념으로서 티나샤키(Tina Sharkey)가 사람들의 관계를 여여주는 서비스를 총칭하는 말로써 이 개념을 사용했다(Tina Sharkey, 2008). 또한 미국 가이드와이어(Guidewire)사의 크리스시플리(Chris Shipley)는 2004년 SNS를 포함한 웹2.0의 관점으로 소셜미디어의 개념을 논의하였다. 이러한 소셜미디어는 매스미디어와 비교하였을 때 다음과 같은 특징을 갖는다

전통적인 미디어와 달리 소셜미디어는 네트워크를 활용한 확산이 가능하고, 신속성과 파급력을 지니며 개인적인 관심 분야를 기초로 하는 기사작성과 함께 기존 저널리즘 뉴스를 필터링 하는 방식으로 저널리즘적인 행동을 하고 있다(이진형, 2012). 특정이슈가 여론에 의해서 정책에 반영되는 과정에 대한 연구를 살펴보면 언론에 보도된 이슈가 대중의 관심을 이끌어내고, 이는 곧 미디어 아젠다에서 공공아젠다로 발전하며, 대중의 관심이 그 이슈에 계속 집중되면 그 이슈는 공공아젠다에서 정부아젠다의 위상으로 진화한다(박기목, 2015). 이 때 소셜미디어는 사회적 이슈에 대한 대중의 관심을 파악하는 도구로 사용될 수 있다. 소셜미디어를 통한 대중의 동향을 파악한 연구로 이수진·전유나(2016)는 소셜미디어에서 거론되는 관광 관련 빅데이터를 분석하여 대중의 인식을 파악하고, 이를 통해 관광분야에서의 소셜미디어 빅데이터 활용방안 모색에 방향을 제시

하였으며, 홍지숙·오익근(2016)은 항공사의 특정 사건을 전후로 대중들의 항공사에 대한 이미지 변화에 대해서 네이버 뉴스와 블로그와 같은 소셜미디어 데이터 분석을 통해 연구하였다. 오세종·김치호(2016)는 오피니언 마이닝 기법을 활용하여 웹드라마 분석을 통해 웹드라마에 아이돌그룹 멤버의 출현과 포털사이트의 편성회사와의 연관성에 대하여 연구하였다. 공공 혹은 정부 정책 관련하여 진행한 소셜미디어 데이터 분석 연구로 배정환 외(2013)는 2012년 한국대선 기간에 대중이 생산한 트위터 문서를 활용하여 대선과 관련된 이슈의 추이를 분석하였으며, 2014년에는 트위터 데이터 분석을 기반으로하여 트위터 상의 일별 이슈들을 분석하는 연구를 진행하였다(배정환 외, 2014). 또한, 정부 3.0 시행에 따라 공공기관들은 블로그, 트위터, 페이스북 등의 소셜미디어를 통해 정보 공유 및 소통을 위한 노력을 기울이고 있으며, 공공 정보를 개방·공유하고, 부처 간 칸막이를 없애고 소통·협력하는 것을 목표 중의 하나로 삼고 있다. 따라서 새로운 정책을 수립하거나, 특정 정책에 대한 평가의 과정에서의 여론 파악을 위해서 소셜미디어의 저널리즘적 기능의 활용 가능성을 탐색할 필요가 있다. 한편, 소셜미디어는 대중과 직접적인 소통이 가능한 수단이라는 관점에서 정책 의제를 수립하고, 정책을 집행하며, 정책을 평가하는 등의 정책과정 단계에서 활용이 가능하다. 또한, 정책 의제 수립 측면에서는 소셜미디어를 통해 다양하게 제시되는 의견에 대한 수집, 추이 분석을 통해 의제 발굴

〈표 1〉 매스미디어와 소셜미디어 비교

구분	특성
매스미디어	<ul style="list-style-type: none"> • 주요 언론사가 뉴스, 정보, 엔터테인먼트 정보를 생산 • 불특정 다수, 수동적 독자층에게 공적, 일방적, 간접적으로 전달
소셜미디어	<ul style="list-style-type: none"> • 웹2.0 정신(참여·공유·개방) 기반으로 누구나 생산(다수생산)자가 되며, 네트워크 중심 쌍방향 소통이 이루어짐

출처 : 삼성경제연구소(2010)

에 반영될 수 있다(박한우, 2012).

본 연구의 목적은 소셜미디어 기반 데이터 분석을 통해 특정 교육정책에 대한 여론의 동향을 파악하는 것이다. 이를 위해 주요 교육정책관련 키워드를 선정하고, 네이버 블로그, 트위터와 같은 소셜미디어 서비스를 대상으로 키워드를 포함하는 문서를 수집하였다. 수집한 문서를 데이터 전처리를 통해 토큰화 하고, 오픈한글 사전으로부터 구축한 감성사전을 활용하여, 해당 문서의 교육정책 키워드에 대한 감정점수를 산정하였다. 그리고 마지막으로 극성분류 기준을 반영하여 최종적으로 문서에 대한 극성을 분류하였다.

II. 선행연구

1. 소셜미디어

1) 소셜미디어의 출현

소셜미디어가 나타나기 전인 웹1.0의 시기에는 대형 미디어 매체들이 생산한 콘텐츠를 일방적으로 제공하고, 사용자들은 미디어 매체들이 제공하는 서비스 경로를 통하여 콘텐츠를 단순히 소비하는 행태였다. 반면에 소셜미디어의 기술기반인 웹2.0 환경에서는 단순 소비자였던 사용자들이 다양한 플랫폼을 활용하여 콘텐츠를 창작하고, 유통하며, 소비 및 공유를 하게 되었다(허정 외 2016). 또한 스마트폰의 확산으로 개인들은

언제 어디서나 트위터나 페이스북 등에 접근할 수 있게 되면서 내가 아는 정보나 일상을 보다 쉽게 자주 공유할 수 있게 되었다(김형지외 2012). 이러한 웹1.0에서 웹2.0으로의 진화는 소셜미디어를 통한 일반 대중의 사회적 참여 확대에 큰 기여를 했다고 볼 수 있다. 웹 1.0과 웹2.0을 비교하면 다음과 같다.

〈표 2〉에서 알 수 있듯이, 웹2.0이 기존의 웹1.0과 가장 차별화되는 것은 참여와 개방 그리고 공유 중심이라는 것이고, 또 집단적 지성을 이용한 다양성을 추구할 수 있다는 특징이다. 이 웹2.0 기술 환경은 소셜미디어의 기반이 되어 개인들로 구성된 커뮤니티가 온라인상에서 정보를 공유하고 지식과 의견을 나누는 활동을 가능하게 하며, 특정 이슈에 대한 여론형성 기능의 일부를 담당하는 것이 가능하게 만들어 주었다. 이런 이유로 많은 기업과 기관을 중심으로 기업과 상품 브랜드 및 기관에 대한 여론 동향을 파악하고 의사결정을 지원하기 위한 소셜미디어 분석이 활발하게 이루어지고 있다(허정 외, 2016).

2) 소셜미디어 개념 및 특성

소셜미디어는 사회를 구성하는 개인들이 경험과 생각, 특정 주제에 대한 의견 등을 서로 공유하기 위해 사용하는 플랫폼과 도구를 말한다. Boyd & Ellison (2008)은 SNS(Social Networking Service)를 개인이 공적 신상정보를 만들거나 자신과 인맥을 맺은 다른

〈표 2〉 웹1.0과 웹2.0 비교

웹1.0	웹2.0
<ul style="list-style-type: none"> • 포털 위주 • 포털상의 서비스 사용자 변경불가 • 폐쇄적 정보/콘텐츠 • 기술중심 • 정보전달에 대한 효율성 추구 • OS 종속성 높음 • 제작자 : 전문가, 프로그래머, 업체등 • 예시: 홈페이지 등 	<ul style="list-style-type: none"> • 플랫폼 위주 • 다양한 서비스 이용 수단으로서 웹활용 • 정보/콘텐츠 공유 및 개방 • 참여, 공유 중심 • 집단적 지성을 활용한 다양성 추구 • 브라우저 : 웹접속 가능 모든 프로그램 • 제작자 : 개인중심의 네트워크 사용자 • 예시 : Wikipedia, 블로그 등

출처 : 「정보통신방송정책」, 25(9). 한국정보통신정책진흥원

이용자들의 리스트를 분류하고 다른 사람들의 계정에 만들어진 인맥 리스트를 볼 수 있게 한 웹 기반 서비스라고 규정하고 있다. 소셜미디어의 세부적인 유형으로는 블로그, 소셜네트워크서비스, 커뮤니티, UCC, 마이크로블로그 등이 있으며, 한국에서는 1999년 싸이월드의 등장으로 소셜네트워크 개념이 태동되었고, 미국에서는 2005년에 페이스북과 2006년에 사회적 관계 및 정보 공유를 목적으로 140자 이내 단문 메시지를 주고 받는 트위터가 등장하였다(김형지 외, 2012). 소셜미디어를 통해 생산 및 공유되는 저작물의 형태는 텍스트, 이미지, 오디오, 비디오 등이며 블로그, 소셜네트워크, 인스턴트 메시지 보드, 팟캐스트, 위키, 사용자 제작 콘텐츠(UCC)와 같은 서비스를 통하여 제작 및 공유된다(오세중 · 김치호, 2016).

대표적인 소셜미디어 서비스는 소셜네트워크서비스(Social Network Service, SNS), 블로그(Blogs), 위키스(Wikis), 팟캐스트(Podcasts) 등이 있다(이주영, 2013). SNS는 카카오톡, 페이스북과 같이 특정 온라인 사이트를 통해 친구관계를 맺고 이용자가 직접 콘텐츠를 생산하고, 공유하는 서비스로, 친구나 지인과 교류할 수 있는 연결망과 개인의 사적인 정보를 수록할 수 있는 공간을 제공한다는 점에서 오프라인에서 성행되는 관계맺기를 온라인으로 끌어온 인맥형 서비스로 불린다(김유정 · 조수선, 2012). 블로그는 Web(웹)과 Log(일기)의 합성어로 기존의 개인 홈페이지보다 관리가 용이하여 많은 이용자들이 홈페이지를 대체하여 사용하고 있으며, 온라인 백과사전인 위키피디아로 대표되는 위키스는 집단지성을 활용하는 대표적인 사례로 이용자가 온라인을 통하여 백과사전의 내용을 직접 편집할 수 있도록 웹페이지를 구성하여 웹상에 콘텐츠를 직접 추가하거나 정보 편집이 가능한 서비스이다. 팟캐스트는 웹을 통해 사용자들이 인터넷 라디오 방송을 오디오 파일로 구독할 수 있도록 하는 서비스를 의미하며 2004년 말부터 활성화 되고 있는 서비스이다. 소셜미디어의 주요 특성은 웹2.0 환경을 기반으로 하는 참여, 개방, 대화, 연결을 들 수 있다. 특정한 주

제에 대하여 공감대를 공유한 사람들이 자발적으로 서로 의견을 교환하고, 지식을 공유하며, 이러한 참여와 피드백 활동 투표에 활용할 수도 있다. 또한 댓글과 메시지 등을 통한 '쌍방향 대화'를 지향하며, 동일 관심사에 대하여 소셜미디어 이용자들이 온라인상으로 '커뮤니티'를 이루고 이를 중심으로 공통 관심 주제에 대하여 소통한다. 온라인 기반의 대부분의 소셜미디어는 다양한 미디어의 조합과 연결서비스를 통해 네트워크를 구축한다(이주영, 2013). 다양한 미디어의 조합을 통해서 다양한 형태의 융합 콘텐츠를 유통한다. 여기서 융합(Convergence)은 디지털 기술의 발전으로 개별 미디어영역이 서로 수렴되고 고유한 서비스 영역의 경계가 모호해지는 변화를 의미한다(이명진 · 박현주, 2012). 이러한 융합 기술로 생산된 콘텐츠의 생산과 공유 및 가치부여는 온라인 기반의 새로운 마켓을 형성하여 산업 발전에도 기여를 하고 있다.

2. 오피니언 마이닝

1) 오피니언 마이닝 개념

스마트 디바이스의 발달 등으로 소셜미디어에는 다양한 변화가 생겼다. 특히 SNS가 급속히 성장하고 사용자가 급증함에 따라 사람들의 개인적인 의견들이 다양하게 생산되고 있으며, 그 데이터양과 속도가 점차 가속화되고 있다. 반면에, 이용자들이 방대한 데이터로부터 필요한 정보를 찾아내는 과정은 더욱 복잡하고 어려워지고 있다. 소셜미디어가 생산하는 대량의 데이터로부터 사용자가 필요로 하는 정보를 빠르게 찾아내고, 의미있는 정보를 유출해내는 데이터 분석기술도 지속적으로 발달하고 있다. 강만모 외(2012)는 데이터 분석기술에는 텍스트마이닝(Text mining), 평판분석(Opinion Mining), 소셜네트워크분석(Social Network Analytics), 클러스터 분석(Cluster Analysis)이 있다고 하였으며, 그 중에서 텍스트마이닝의 관련 분야인 오피니언 마이닝은 소셜미디어 등으로부터 수집한 정형 혹은 비정형 텍스트에 대하여 긍정, 부정, 중립의 선호도를 판

별하는 기술이라고 하였다. 서민송·유환희(2017)는 오피니언 마이닝이 감성분석을 의미하고, 사람들의 태도, 의견, 성향과 같은 데이터를 활용하여 특정 주제에 대해 긍정적 혹은 부정적인지를 분류하는 기술이라고 하였다. 즉, 오피니언 마이닝은 문서의 주제가 무엇인지 찾아내는 것 보다는, 그 문서의 저자가 주제에 대하여 가지고 있는 감정을 분석한다. 오피니언 마이닝의 개념은 다양한 연구에 활용되고 있는데, 박경미외(2011)는 오피니언 마이닝을 활용하여 어떤 사안에 대한 여론 혹은 대중의 관심이 시간이 지남에 따라 변하는 특징을 활용하여 맛집이나 휴대폰 등의 리뷰 변화에 대하여 분석하였고, 정은희·이병관(2015)은 오피니언 마이닝은 감성 사전을 구축하여 분석에 사용하는 반면, 텍스트 마이닝은 입력 데이터의 패턴에 대한 학습을 통해 다른 데이터를 예측하는 차이점이 있다고 하였다.

2) 오피니언 마이닝 단계

오피니언 마이닝은 기본적으로 문서가 긍정, 부정, 또는 중립 중 어떤 견해를 갖고 있는지 판별하는 일련의 과정이라 볼 수 있으며, 분석은 각 문서 최소 단위인 어휘의 감성 극성(Sentiment Polarity)에 기반하여 이루어진다(김승우·김남규, 2014). 오피니언 마이닝 기법은 기본적으로 다음 세 가지 절차로 구성된다.

첫째, 분석은 문서 단위로 이루어지며, 문서로부터 긍정이나 부정을 표현하는 어휘와 문구를 추출하고, 긍

정과 부정을 표현하는 기준단어로 구성된 감성사전을 구축한다.

둘째, 오피니언으로 구성된 문장을 인식하고 세부 평가요소를 분류한다. 이때 첫 번째 단계에서 구축되어 분류된 감성관련 어휘 정보를 사용한다. 또한, 저자의 감정이 긍정적인지 부정적인지 문장 단위로 분류하기 위해서 여러 가지 규칙기반 혹은 통계기반 방법론을 동시에 적용하기도 한다.

셋째, 텍스트의 긍정과 부정 표현의 출현 빈도수를 기반으로 극성분류기준을 만들고 그에 따라 문서의 극성을 분류한다.

오피니언 마이닝의 주요 기법으로는 Linguistic Approach, SVMs(Support Vector Machines), Lexicon-based Approach 등이 있다.

- Linguistic Approach: 극성을 판별할 때 텍스트의 문법적인 구조를 파악하여 활용한다. 주로 Lexicon-based Approach 방식과 함께 사용하며, 문맥(context)을 파악하여 극성을 판별한다.
- SVMs: 텍스트의 긍정/부정 의견을 분류할 때 사전에 분류된 학습 데이터(Training Sets)를 활용하는 방식이다.
- Lexicon-based Approach: 사전에 정의한 긍정 표현과 부정표현 단어세트(Bag of Words)를 이용하여 텍스트에 포함된 단어 출현 빈도를 활용하여

〈표 3〉 최종 선정 교육 정책 및 키워드

소프트웨어 교육	<ul style="list-style-type: none"> · 소프트웨어 교육(SW 교육) · 소프트웨어 의무(SW 의무) · 코딩 교육 	<ul style="list-style-type: none"> · Computer science(컴퓨터과학) · Computational Thinking(컴퓨터적 사고)
스마트 교육	<ul style="list-style-type: none"> · 스마트 교육(smart 교육) · 디지털교과서(digital textbook, 전자교과서, e-textbook) 	<ul style="list-style-type: none"> · 스마트러닝(smart learning) · 모바일러닝(mobile learning) · 이러닝(e-learning)
일반 정책	<ul style="list-style-type: none"> · 자유학기제 · 융복합교육 	<ul style="list-style-type: none"> · 창의, 인성교육(창의교육, 인성교육)

긍정과 부정을 판별하는 방법으로 LIWC(Linguistic Inquiry and Word Count)나 POMS (Profile of Mood States) 같은 사전을 이용할 수 있는데 한국어에는 적용되지 않는다.

Ⅲ. 연구방법

1. 교육정책 키워드 선정

소셜미디어 기반 빅데이터 분석을 위한 교육 정책관련 키워드를 선정하기 위해 본 연구에서는 기초자료로 2015 교육정보화백서를 중심으로 2016 교육부 업무계획, 교육유관기관 교육정책연구협의회 자료를 활용하였다. 교육정책은 지능정보사회 도래에 따라 관심을 받고 있는 스마트 교육정책과 소프트웨어 교육 정책, 그리고, 일반정책 그룹으로 구분하여 후보 키워드들을 추출하였고, 교육관련 분야 박사 9명으로 구성된 전문가협의회 검토를 거쳐 데이터 분석 키워드를 최종적으로 선정하였다.

2. 소셜미디어 데이터 수집

1) 소셜미디어 데이터 수집대상 선정

교육정책에 대한 소셜미디어 기반 여론의 감정분석을 실시하기 위하여 여러 소셜미디어 서비스 중에서 네이버 블로그와 트위터 두 서비스를 대상으로 데이터를 수집하였다. 다음 블로그는 다음에서 제공하는 API 서비스를 사용하는 응용프로그램에 대해 제한적으로 무료서비스를 제공하고, API 트랙픽 쿼터 정책이 엄격하게 적용되는 등 제약조건이 많아 연구를 위한 데이터 수집에 적절하지 않으며, 페이스북과 같은 소셜네트워크서비스는 데이터 수집을 위해서 개인 혹은 그룹 내 구성원들의 동의를 얻어야 하는 문제가 있어 데이터 수집이 비교적 용이한 네이버 블로그와 트위터를 대상으로 선정하였다.

(1) 네이버 블로그

JAVA 프로그래밍 언어를 기반으로 자체 웹크롤러를 개발하여 네이버 블로그의 데이터를 수집하였으며, 세부 내용은 다음과 같다.

- 입력: URL 클래스의 OpenStream; Buffered InputStream
- 출력: FileOutputStream, mySql DB

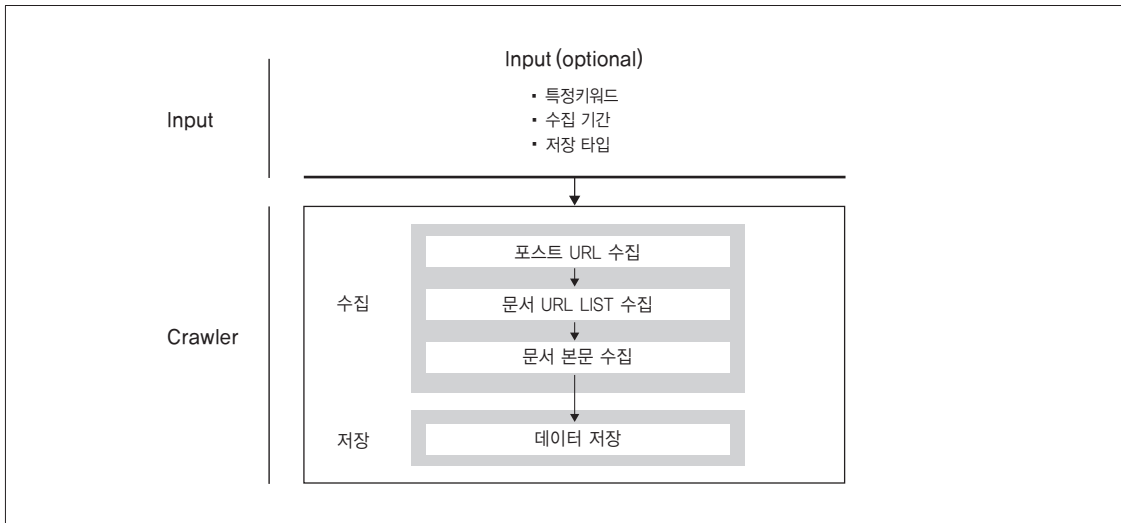
본 연구에서 개발한 웹 크롤러는 아래와 같은 기능을 가지고 있다.

- 실시간 혹은 주기적으로 문서 수집 기능
- 특정 키워드별 문서 수집
- 다양한 값의 입력 변수 설정을 통해 사용자의 조건에 부합되는 데이터 수집
- 수집된 데이터는 연결된 데이터베이스에 저장

크롤러의 구조는 <그림 1>과 같다. 총 3단계를 거쳐 네이버 블로그의 문서 수집을 진행하였으며 키워드 및 수집 기간 등 다양한 변수 설정을 통해 원하는 조건에 맞는 데이터를 수집하였다. 특히 키워드의 경우에는 일치검색 조건을 이용한 “키워드”로 검색하도록 하여 수집하였다.

(2) 트위터

트위터는 2006년 3월 서비스를 시작한 마이크로블로그 웹 사이트이다. 트위터는 140자로 제한된 콘텐츠 제작만을 허용한다. 이러한 제한은 콘텐츠 내용을 다른 소셜네트워크서비스보다 간결하고 명확하게 전해주는 특징이 있다고 할 수 있다. 트위터는 다수의 기능을 OpenAPI로 제공하고 있으며, 다양한 업체들이 OpenAPI를 통해서 트위터 데이터를 분석하거나 검색하며, 클라이언트 애플리케이션을 개발하고 있다. 본 연구에서도 트위터 데이터를 수집하기 위하여 트위터가 제공하는 OpenAPI를 사용하였다. 현재 java로 된 트위터 라이브러리로는 Twitter4J, java-twitter,



〈그림 1〉 웹 크롤러 아키텍처

jtwitter, Twitter Client 등이 있으며, 이 중에서 Twitter4J를 선택하였다. 또한 트위터는 REST API와 Streaming API 두 가지 방식을 제공하는데, REST API는 특정 단어의 검색 결과, 특정 사용자 정보와 같은 특정 작업을 수행하는데 적절하고 Streaming API는 트위터 데이터를 실시간으로 받아볼 때 사용한다. 본 연구에서는 REST API의 Search API를 사용하였으며, 트위터의 외부에서의 인증을 위해서 OAuth인증 방식을 사용하고 있다. OAuth인증 방식은 Service가 인증정보를 가지고 있는 User가 Service의 인증정보로 Consumer사이트를 이용하려는 것이다. 그 인증방식의 흐름은 다음과 같다.

- 사용자 홈페이지에서 서비스제공자에게 Request Token 요청
- 서비스 제공자가 Request Token 발급
- 발급받은 Request Token과 함께 사용자를 서비스제공자에게 이동
- 서비스제공자가 사용자인증하고 다시 사용자 홈페이지로 이동
- 사용자 홈페이지에서 Access Token 요청.

- 서비스제공자가 Access Token 발급.
- 발급받은 Access Token을 이용하여 서비스제공자 자원 사용. 트위터 API는 무분별한 트래픽 증가를 방지하기 위하여 다음과 같은 제약사항을 둔다.
- 익명의 이용자가 API로 요청 시 시간당 150회의 요청으로 제한하고 있으며 ID를 인증한 경우 15분당 180회의 요청이 가능하다.
- 1회 요청 시 최대 100개의 트위터 자료를 가져올 수 있다.
- 트위터로부터 제공되는 API로부터 최근 10일간 관련된 키워드가 들어있는 트위터 문서들을 수집할 수 있으며 10일 이내에는 일 단위로 임의로 날짜를 선택하여 수집할 수 있다. 트위터 API를 통하여 사용자, 작성시간, 트위터 콘텐츠 지리적 위치, 해쉬 태그, 리트윗 관련정보 등을 가져올 수 있다. 본 연구에서는 트위터 API 수집기를 이용하여 수집한 데이터 중에서 사용자, 작성 시간, 트위터 콘텐츠 등 속성들을 추출하여 연구에서 사용하였다.

2) 소셜미디어 데이터 수집

네이버 블로그와 트위터를 대상으로 데이터를 수집

〈표 4〉 소셜 미디어 데이터 수집 기간

소셜미디어	수집 기간
네이버 블로그	2016년 5월 1일~ 2016년 8월 31일
트위터	2016년 5월 8일~ 2016년 8월 31일

〈표 5〉 소셜미디어 데이터 저장 테이블 스키마

항목명	항목 설명	데이터 형	예
docNo	문서번호	varChar	N0000000001(NAVER Blog); A0000000001(Naver News); T0000000001(twitter)
keyword	키워드	varChar	소프트웨어 교육, SW 교육
date	작성일	int	20160705
time	작성시간	int	121800(12시 18분 00초)
publisher	작성자	varChar	tinyedel
URL	주소	varChar	http://blog.naver.com/PostView_nhn?blogId=tinyedel&logNo=220754024141&beginTime=0&jumpingVid=&from=section&redirect=Log&widgetTypeCall=true
title	제목	varChar	SW교육 전문가 양성과정 이수완료!
content	내용	longText	생략 (문서 본문 내용)

〈표 6〉 문장 토큰화 예시

예시문장	학술정보란 학문과 기술에 관한 자료를 실제 문제에 도움이 되도록 정리한 지식이나 자료를 말한다.
토큰	학술정보란, 학문과, 기술에, 관한, 자료를, 실제, 문제에, 도움이, 되도록, 정리한, 지식이나, 자료를, 말한다.

한 기간은 〈표 4〉와 같다.

수집한 데이터는 MySQL을 이용하여 데이터베이스에 테이블 형태로 저장된다. 〈표 5〉는 소셜미디어 데이터를 데이터베이스에 저장할 때 사용되는 테이블 스키마에 대한 설명을 나타낸다. 항목명은 테이블의 각 항목에 대한 이름을 나타내고, 항목 설명은 이에 대한 설명, 데이터 형은 각 항목의 데이터 형식 그리고 예는 각 항목에 어떠한 값이 들어가는지에 대한 예시를 나타낸다.

3. 감정분석

1) 토큰화

문서의 극성은 문서에 있는 단어 단위로 분석을 진행하고, 그 누적된 결과를 활용하여 문서 단위로 극성을 판단한다. 본 연구에서는 수집한 모든 문서를 대상으로 공백, 줄바꿈과 문장부호 기준으로 단어 단위의 토큰으로 분리하는 작업을 진행한다. 〈표 6〉은 문장을 토큰화한 결과 예시이다. 아래의 예시문장은 〈학술정보란, 학문과, 기술에, 관한, 자료를, 실제, 문제에, 도움이, 되도록, 정리한, 지식이나, 자료를, 말한다.〉 등 13개의 토큰으로 분리할 수 있다.

〈표 7〉 감정자질 추출예시

예시문장	준비가 미흡하다는 의견도 있고 가르칠 교사가 턱없이 부족하다는 의견도 있고 하지만, 그래도 이전 스마트교육 추진보다는 방향성 면에서 괜찮다고 생각합니다.
감정자질	미흡하다는 의견도, 턱없이 부족하다는, 부족하다는 의견도, 괜찮다고 생각합니다.

2) 감정자질 추출

본 연구에서는 기존 연구를 통해 구축된 오픈한글 API를 이용하여 감성 사전 데이터베이스로부터 감정자질을 추출한다. 오픈한글은 한글 오픈데이터 플랫폼으로써 한글의 자연어 처리뿐만 아니라 감정 분석에 관련한 API를 제공하는 서비스이다. 이는 집단지성을 활용한 감성이 사전을 구축하고 이를 활용한 감정분석 서비스도 제공한다(안정국·김희웅, 2015). 감정자질은 문서 내 감정의 특징을 표현할 수 있는 어휘들이며, 본 연구에서는 “형용사+명사”, “형용사+동사”, “부사+형용사”, “명사+동사” 등 5가지 조합으로 연속되어 사용된 어휘를 말한다. 〈표 7〉은 감정자질을 추출한 예시이다.

본 연구에서는 수집한 문서들을 대상으로 위와 같은 감정 자질 추출의 절차를 거쳐서 감정 자질을 추출하였고, 적어도 한 개 이상의 감정자질이 추출된 문서들에 대해서 감정분석 작업을 진행하였다.

3) 감정점수 계산

추출한 감정 자질들에 대한 감정점수를 계산하기 위해 기존 감정분석 연구에서 일반적으로 적용되는 PMI-IR(Pointwise Mutual Information-Information Retrieval)(Turney, 2002)을 활용하여, 본 연구의 취지에 맞게 수식을 설계하였다. PMI는 PMI-IR 알고리즘은 문구(Phrase)의 의미적 지향성을 측정하기 위하여 사용된다. PMI-IR은 문구와 의미를 가진 기준단어가 함께 출현한 문서 검색수차이를 통해서 의미적 지향성을 판단하는 방법이다. 문구의 의미적 지향성은 그 차이가 클수록 강하다고 할 수 있다(Turney, 2002). 본 연구에서는 각 문서에서 추출한 감정 자질 문구의 감정

점수를 각 감정자질이 긍정 단어와 함께 쓰였을 때와 부정 단어와 함께 쓰였을 때의 문서 검색결과수의 비로 표시 하고, 그 값의 합의 평균을 감정점수로 산출하였으며, 감정점수 수식은 다음과 같다.

$$SS(phrase) = \frac{\sum_i^n \sum_j^m \log_2 \left(\frac{hits(phrase \text{ and } P_i)}{hits(phrase \text{ and } N_j)} \right)}{n \times m}$$

- $SS(phrase)$ phrase의 감정 점수
- P_i i번째 긍정 기준 단어 $i=1, \dots, n$
- N_j j번째 부정 기준 단어 $j=1, \dots, m$

- $hits(phrase \& P_i)$ phrase와 P_i 가 함께 출현한 문서 검색 수
- $hits(phrase \& N_j)$ phrase와 N_j 가 함께 출현한 문서 검색 수

이 수식을 적용하여 감정 점수를 계산하기 위해서는 기준 단어 집합을 정의하는 것이 중요하며, 기준 단어의 검색수가 너무 많거나 적을 경우 감정자질의 극성을 판단하는데 있어 왜곡이 생길 수 있고, 이로 인해 정확한 극성 구분이 불가능 할 수 있다. 따라서, 본 연구에서는 이러한 왜곡을 줄이고, 잘 못 해석 될 수 있는 가능성을 최소한으로 줄이기 위하여 오픈한글 API를 통해 얻은 감정 점수와 감정을 기준으로 감정점수가 100%인 긍정어, 부정어들 중에서 출현빈도가 가장 높은 상위 10개 단어를 선정하여 분석에 사용하였다. 〈표 8〉은 감정점수 계산에 사용되는 기준단어인 긍정과 부정어 단어 집합을 나타낸다. 본 연구에서는 특정 감정 자질의 감정점수를 계산하기 위해 아래 표에 나타난 극

〈표 8〉 긍정과 부정 기준 단어 집합

	긍정단어	부정단어
1	있다	없다
2	좋다	못하다
3	많다	부족하다
4	사랑스럽다	아니다
5	가능하다	불편하다
6	감사하다	잘못하다
7	좋아하다	위험하다
8	쉽다	늦다
9	재미있다	실패하다
10	만족하다	나쁘다

phrase가 기준 단어들과 함께 출현한 hit수

phrase	phrase&P ₁		phrase&N ₁	
	검색어	hit	검색어	hit
너무 재미있고	"너무 재미있고"&"있다"	1,273	"너무 재미있고"&"많다"	195
	"너무 재미있고"&"좋다"	29	"너무 재미있고"&"못하다"	284
	"너무 재미있고"&"많다"	236	"너무 재미있고"&"부족하다"	314
	"너무 재미있고"&"사랑스럽다"	14	"너무 재미있고"&"아니다"	43
	"너무 재미있고"&"가능하다"	275	"너무 재미있고"&"불편하다"	4
	"너무 재미있고"&"감사하다"	105	"너무 재미있고"&"잘못하다"	108
	"너무 재미있고"&"좋아하다"	306	"너무 재미있고"&"위험하다"	62
	"너무 재미있고"&"쉽다"	12	"너무 재미있고"&"늦다"	788
	"너무 재미있고"&"재미있다"	893	"너무 재미있고"&"실패하다"	607
	"너무 재미있고"&"만족하다"	116	"너무 재미있고"&"나쁘다"	8

phrase의 감정 점수 계산

$$\begin{aligned}
 SS(\text{"너무 재미있고"}) &= \text{average} \left\{ \log_2 \left(\frac{\text{hits}(\text{"너무 재미있고"} \& \text{"있다"})}{\text{hits}(\text{"너무 재미있고"} \& \text{"없다"})} \right), \dots, \log_2 \left(\frac{\text{hits}(\text{"너무 재미있고"} \& \text{"만족하다"})}{\text{hits}(\text{"너무 재미있고"} \& \text{"나쁘다"})} \right) \right\} \\
 &= \text{average} \left\{ \log_2 \left(\frac{1273}{195} \right), \dots, \log_2 \left(\frac{116}{8} \right) \right\} \\
 &= \text{average} \{ 2.706686, \dots, 3.857981 \} \\
 &= 0.412
 \end{aligned}$$

〈그림 2〉 감정 점수 계산 예시

성단어와 함께 등장한 문서 수를 이용하였다.

감정자질이 긍정적일수록 감정 점수 값이 크고 반대로 부정적일수록 감정 점수 값은 작은 값이 나온다.

아래 〈그림 2〉은 "너무 재미있고"라는 부사와 형용사의 조합으로 구성된 감정자질의 감정 점수를 계산하

는 예시이다. 먼저 네이버 블로그에서 감정 자질과 긍정, 부정 단어 집합들이 함께 쓰인 블로그 건수를 수집한다. 이러한 검색건수를 바탕으로 위에서 설명한 수식에 대입하여 감정 점수의 값을 구하면 "너무 재미있고"라는 감정 자질은 예시에서 보여주는 바와 같이 0.412

〈표 9〉 문서의 감정 점수 예시

Phrase	Sentiment Score
쉽게 공부할	0.772
아주 좋은	1.567
다양한 교육	0.433
문서의 감정 점수	0.924

〈표 10〉 주요 키워드별 문서 수집결과

구분	수집결과	비고
소프트웨어 교육 이슈 관련 9개 키워드	전체 1,330건 문서수집완료. 코딩교육 키워드가 총 408건으로 전체 수집문서의 30.6%를 차지함	
스마트교육 이슈 관련 12개 키워드	전체 911건 문서수집완료. 이러닝 키워드가 총 448건으로 전체 수집문서의 49.1%를 차지함	
일반정책 이슈 관련 5개 키워드	전체 2,268건 문서수집완료. 인성교육 키워드가 총 1,153건으로 전체 수집문서의 50.8% 차지함	

이라는 감정 점수를 갖는 것을 알 수 있다.

하나의 문서의 감정 점수는 해당 문서에서 추출된 감정자질들의 감정 점수의 평균으로 표시할 수 있다. 예를 들어서 doc_001에 포함되는 감정자질이 총 3개이고 아래와 같을 때 doc_001의 감정 점수는 각 감정자질의 감정 점수의 평균으로 계산 할 수 있다.

IV. 연구결과

1. 문서 극성 분류

본 연구에서 선정한 키워드를 기반으로 네이버 블로그와 트위터를 대상으로 수집한 문서의 기초 통계량은 〈표 10〉과 같다.

네이버 블로그와 트위터로부터 수집한 모든 문서들로부터 토큰화, 감정자질 추출을 거쳐, 감정점수를 계산하고, 감정자질이 추출 되지 않은 문서를 제외한 문서들의 감정 점수의 분포를 그려보면 아래 〈그림 3〉과 같이 나타낼 수 있다.

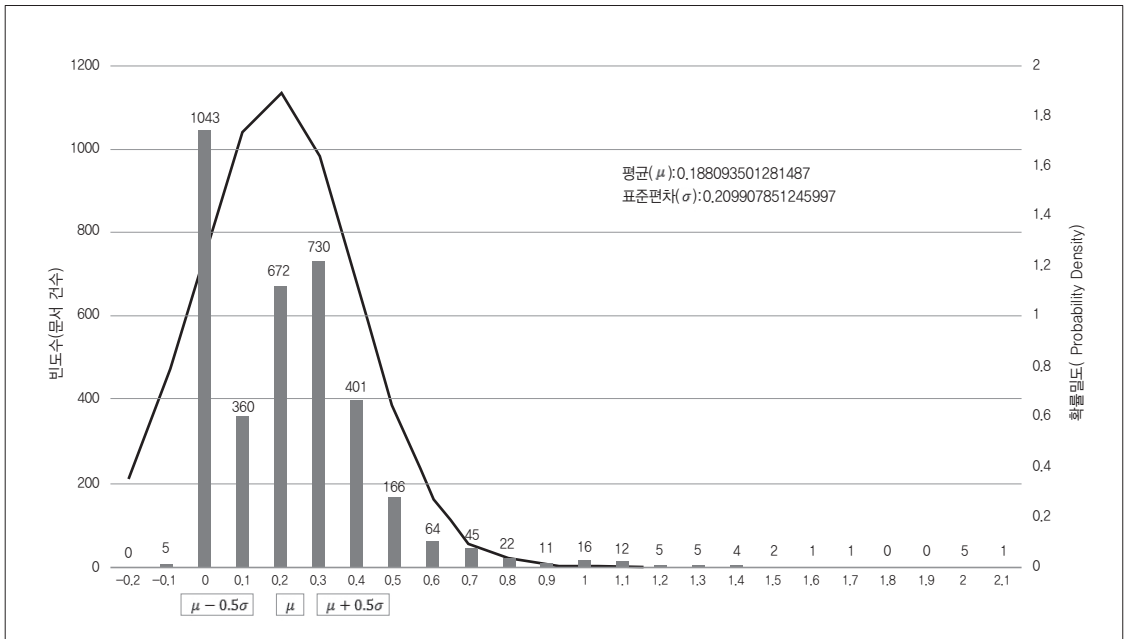
〈그림 3〉은 전체 문서의 감정점수들의 평균값(μ)과 표준편차(σ)를 구한 후, 감정점수 구간(최소값, 최대값 사이 0.1 구간 단위)에 따른 히스토그램과, 정규분포 ($N(\mu, \sigma)$) 확률밀도함수(Probability Density Function)를 그래프로 나타낸 것이다.

위와 같은 히스토그램을 바탕으로 본 연구에서는 분석 대상 문서의 감정을 긍정, 중립, 부정 3가지로 분류하였으며, 이 때 극성분류의 기준은 평균값(μ)과 표준편차(σ)를 이용하여($-\text{inf}, \mu-0.5*\sigma$), ($\mu-0.5*\sigma, \mu+0.5*\sigma$), ($\mu+0.5*\sigma, \text{inf}$)으로 계산하였다.

$$\text{polarity} = \begin{cases} \text{긍정} & \text{where } SS(\text{Doc}) \geq 0.293047 \\ \text{중립} & \text{where } 0.293047 > SS(\text{Doc}) \geq 0.08314 \\ \text{부정} & \text{where } SS(\text{Doc}) < 0.08314 \end{cases}$$

이러한 극성 분류 기준을 이용해 전체 문서에 대하여 극성 분류를 진행하면 〈표 10〉과 같은 결과를 얻는다.

전체 4,509건 중에서 감정자질이 추출되지 않은 문서 938건이 있었으며 이런 문서들은 감정 분류가 불가



〈그림 3〉 감정점수 히스토그램

능한 것으로 분류되었으며 나머지 3,571건에 대한 감정 분석 결과는 다음과 같다.

소프트웨어교육, 컴퓨터적 사고, 코딩교육 등의 키워드로 수집된 소프트웨어 교육 정책 문서들은 긍정이 319건으로 32.3%, 중립이 381건으로 38.5%, 부정이 289건으로 29.2%를 나타내어 대체적으로 중립에서 긍정적인 감정으로 나타났다. 디지털교과서, 이러닝, 스마트러닝 등의 키워드로 수집된 스마트교육관련 문서들에 대해서는 긍정이 205건으로 28.6%, 중립이 261건으로 36.4%, 부정이 252건으로 35%를 나타내어 중립에서 부정적인 감정으로 나타났다. 또한 자유학기제나 융·복합 교육 등의 키워드로 수집된 일반정책 문서들은 긍정이 999건으로 25.5%, 중립이 35.6%, 부정이 38.9%를 나타내어 부정적인 감정이 가장 크게 나타났다.

V. 결론

본 연구에서는 교육정책을 수립하기 위한 의제 혹은

어젠다를 설정할 때 대중의 여론을 반영하기 위한 방법으로 블로그와 트위터와 같은 소셜미디어서비스기반의 데이터에 대하여 감정분석을 실시하였다. 이를 위하여 감정분류기법인 PMI-IR기법에 확률밀도함수 개념을 적용하여 연속선상의 감정 극성분류기준을 만들고 수집된 문서를 분류하였다. 〈표 11〉에서는 이러한 3가지 정책에 대한 감정분석 결과를 파악할 수 있으며, 결과를 분석하면 다음과 같다.

첫 번째, 교육정책 중에서도 정보화 정책이라고 할 수 있는 “소프트웨어교육”과 “스마트교육”정책에 비해서 “일반정책”에 대한 감정분석 결과가 비교적 부정적으로 나타났으며, 이는 일반 대중에게 방송이나 뉴스 등을 통해서 노출되는 정책 키워드의 빈도수 차이거나 일반 교육정책과 정보화정책의 차이 때문에 생기는 결과일 수 있다. 두 번째, 3가지 정책 모두 중립인 감정의 비율이 30%이상을 나타내고 있는 것은 교육관련 정책에 대해서는 블로그나 트위터를 통해서 감정의 극성이 치우친 의견이 많이 나타나지 않는다는 것을 보여준

〈표 11〉 문서의 극성 분류 결과

극성분류	문서 건수			
	소프트웨어교육	스마트 교육	일반 정책	합계
긍정	319	205	475	999
중립	381	261	663	1,267
부정	289	252	726	1,305
합계	989	718	1,864	3,571

다. 마지막으로 전체 수집한 4,509건 중에서 감정자질이 추출되지 않은 문서가 938건으로 약 20%나 된다는 것은 우리나라의 소셜미디어가 생산하는 문서들에 저자의 감정 자질 자체가 발견되지 않는 것으로서 단순히 뉴스를 링크한다거나, 특정 주제를 중심으로 사실적인 내용만 기술이 되어 있는 경우가 될 것이다. 이는 특정 주제와 연관된 키워드를 포함한 문서만을 수집하였는데도 나타나는 결과라고 볼 때 우리나라의 소셜미디어가 가지는 특징이라고 할 수 있다.

본 연구를 통해 웹 2.0시대를 뛰어넘어 빅데이터 시대를 맞이하는 현 시점에서 여론형성의 큰 부분을 차지하는 소셜미디어 기반 문서들의 감정분석을 실시하는 실험을 통해 교육정책에 대한 여론 동향을 파악하는 것이 가능함을 보여주었다. 또한 웹 2.0 기술이 발전하면서, 이미 웹 기반의 문서들을 데이터베이스화하여 분석할 수 있는 다양한 기술과 방법론에 대한 연구가 많이 진행되어 왔기 때문에 앞으로 남은 과제는 이러한 방법론들을 통해 소셜미디어를 통해서 대량으로 생산되는 빅데이터를 활용하여 데이터를 분석하는 숙제가 남았다고 할 수 있다. 또한 본 연구는 2016년 5월 ~ 8월까지 약 4개월간 데이터를 수집하여 분석하였지만, 향후에는 데이터 수집을 연단위로 실시하여 정책에 대한 여론감정의 트렌드 분석과 함께 시계월별 비교 등 향후 정책 수립에 기여할 수 있는 연구로 확대 진행해야 할 것이다.

VI. 향후 연구

기존 연구들에서는 특정 주제 인식에 대한 분석과 함께 인식을 강화할 수 있는 방안을 제시하거나, 정책의 제 설정에서의 소셜미디어와 매스미디어의 역할을 비교한 연구 등이 진행되었다. 그러나 소셜미디어가 생산하는 데이터를 기반으로 정책수립 강화방안에 대한 연구는 미미하다고 할 수 있다. 이렇게 본 연구과 같이 여론 탐색으로서의 소셜미디어 데이터 감정분석결과는 정책을 수립하는 과정에서 유용하게 활용될 수 있다. 즉 기존의 정책에 대한 유지 혹은 폐지를 결정할 때, 혹은 새로운 정책을 발굴 할 때 해당 정책에 대한 여론을 분석하여 대중들이 얼마나 긍정적인지 혹은 부정적인지를 분석하고, 해당 정책에 대해 대중들의 성향이 대립되고 있는지 아닌지를 분석하여 정책 결정에 반영할 수 있다. 정책에 관한 일련의 과정은 정책의제설정-정책결정-정책 집행-정책평가 및 환류로 단계별 정리를 할 수 있다. 이와 같은 정책 단계별 정책에 대한 키워드 중심의 소셜미디어 데이터 분석 결과의 내용을 반영할 수 있을 것이다.

또한 향후 연구에서는 감정자질을 추출하는 감성사전을 집단지성기반의 오픈 API를 활용하는 것에서 나아가, 오피니언 마이닝에서 활용되는 기계학습 방법을 활용, 데이터 분류기준을 구축하여 더 정확하고, 확률 높은 극성분류를 하는 시도를 해볼 수 있을 것이다. 또한 이번 연구에서는 오피니언 마이닝 기법을 활용하여

불특정 다수가 생산한 문서에 대한 감정분석만을 실시하였지만, 향후에는 특정한 저자들을 대상으로 감정성향을 분석하여 저자의 감정성향과 문서 극성과의 관계나, 결과적으로 소셜미디어 여론에 끼치는 영향 등 다양한 관점의 소셜미디어 기반 여론분석을 시도해 볼 수 있을 것이다. 나아가서는 정보보호 정책, 국가 정보화 정책 등 다양한 분야에서 이러한 소셜미디어 기반의 여론을 정책에 반영하기 위한 데이터 분석 연구로 확대를 하는 시발점이 될 것으로 판단된다.

■ 참고문헌

- 강만모 · 김상락 · 박상무 (2012). “빅데이터의 분석과 활용.” 『정보과학회지』, 30, 6: 25-32
- 김승우 · 김남규 (2014). “오피니언 분류의 감성사전 활용효과에 대한 연구.” 『지능정보연구』, 20(1): 133-148.
- 김유정 · 조수선 (2012). “미디어 영역에서의 소셜네트워크 서비스 연구동향 분석.” 『정보화정책』, 19(4) : 3-26.
- 김형지 · 최홍규 · 김성태 · 안미선 · 이유민 (2012). “소셜미디어 이용거부자의 혁신저항에 관한 연구.” 『한국언론학보』, 56(4): 439-464.
- 박경미 · 박호선 · 김형곤 · 고희동 (2011). “SNS에서 오피니언 마이닝 연구.” 『정보과학회지』, 29(11): 54-60.
- 박기목 (2015). “언론의 공공의제 설정 및 정책결정 과정과 사례적용에 관한 연구.” 『한국정책학회보』, 24(3): 29~59.
- 박한우 (2012). “저비용 고효율 SNS로 여론 읽기.” 『신문과 방송』, 499.
- 배정환 · 손지은 · 송민 (2013). “텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석.” 『지능정보연구』, 19(3): 141-156.
- 배정환 · 한남기 · 송민 (2014). “토픽 모델링을 이용한 트위터 이슈 트래킹 시스템.” 『지능정보연구』, 20(2): 109-122.
- 서민숙 · 유환희 (2017). “오피니언 마이닝 기법을 이용한 사회적 재난의 시민감성도분석.” 『한국지형공간정보학회지』, 25(1): 37-46.
- 안정국 · 김희웅 (2015). “집단지성을 이용한 한글 감성어 사전 구축.” 『지능정보연구』, 21(2): 49-67.
- 오세종 · 김치호 (2016). “소셜미디어 빅데이터의 텍스트 마이닝과 오피니언 마이닝 기법을 활용한 웹드라마 분석과 제안.” 『한국만화애니메이션학회지』, 285-306.
- 이병진 · 박현주 (2012). “SNS에 관한 사회학 연구의 동향과 쟁점.” 『정보화정책』, 19(2):3-20.
- 이수진 · 전유나 (2016). “소셜미디어 빅데이터 분석을 통한 관광 인식 조사.” 『GR연구논총』, 18(1): 83-109.
- 이종혁 · 김원상 · 박재원 · 최재현 (2014). “오피니언 마이닝을 활용한 블로그의 극성 분류기법.” 『한국디지털콘텐츠학회지』, 15(4): 559-568.
- 이주영 (2013). “소셜 미디어 서비스 현황 및 활용.” 『정보통신방송정책』, 25(9): 45-65. 한국정보통신정책연구원.
- 이진형 (2012). “SNS(Social Network Service)의 확산과 동향.” 『한국방송통신전파저널』, 44. 한국방송통신전파진흥원.
- 정은희 · 이병관 (2015). “오피니언 마이닝 기반 SNS 감성 정보 분석 전략 설계.” 『한국정보전자통신기술학회 논문지』, 8(6): 544.
- 허정 · 이충희 · 오효정 · 윤여찬 · 김현기 · 조요한 · 옥철영 (2016). “소셜 빅데이터 마이닝 기반 이슈 분석보고서 자동 생성.” 『정보처리학회지』, 3(12): 553-564.
- 홍지숙 · 오익근 (2016). “소셜미디어 빅데이터를 활용한 항공사 이미지 변화 분석.” 『관광연구저널』, 30(6): 119-133.
- Boyd, D. M. & Ellison, N. B. (2008). 『Social network sites: Definition, history, and scholarship』. *Journal of Computer-Mediated Communication*, 13, 210-230.
- Turney, Peter D. (2010). 『Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews』. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics