빅데이터 분석과 머신러닝을 활용한 특정 정치인의 견해와 평판에 대한 프로파일링 기술

김민희, 강재은, 최주영, 황채연, 김명주
서울여자대학교 정보보호학과
xwoud@swu.ac.kr

Profile Generation on a Politician' Views and Reputations by using Big Data Analysis and Machine Learning

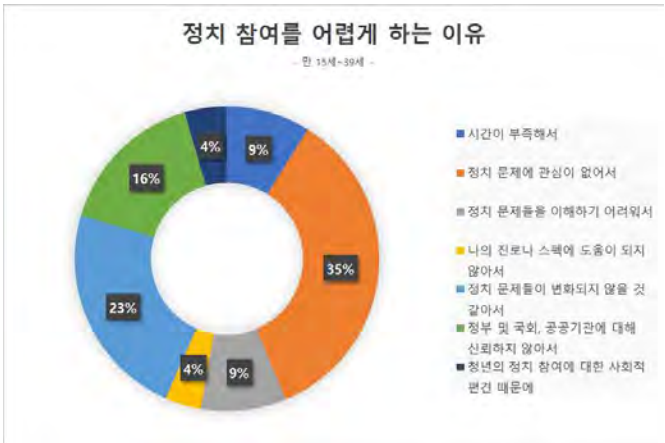
Min-Hee Kim, Jae-Eun Kang, Ju-Yeong Choi, Chae-Yeon Hwang, Myuhng-Joo Kim
Dept. of Information Security, Seoul Women's University

요 약

선거 기간 때마다 유권자들은 어떤 후보자에게 투표권을 행사해야 올바른 선택을 하게 될지 고민하게 되며, 후보자의 선거캠프에서는 후보자에 대한 유권자의 평판에 관심을 가지게 된다. 이러한 고민을 해결하기 위하여 본 논문에서는 TF-IDF 기법과 양방향 LSTM 기계학습모델을 활용해 특정 정치인의 분야별 행보와 여론에 대해 시계열 파악이 가능한 프로파일 보고서를 생성한다. 이를 통해 유권자는 후보자의 정치 철학과 경륜에 대한 이해가 쉬워져 올바른 투표권을 행사할 수 있으며 선거캠프에서는 데이터 기반 평판에 대한 올바른 선거전략을 수립할 수 있게 된다.

1. 연구 배경 및 목적

선거기간이 다가오면 많은 유권자들은 누구에게 투표해야 할지 모르겠고 어떤 후보가 무슨 공약을 세웠는지 관심을 갖지 않으면 알기 어렵다. 또한 이 공약을 제대로 수행할 수 있을지에 대한 신뢰성을 어떤 근거를 바탕으로 받아들여야 하는지 고뇌를 겪게 된다. 통계자료에 따르면 10~30대 연령층의 경우 정치 문제에 관심이 없어서, 정치 문제들을 이해하기 어려워서라는 다양한 의견을 표출하기도 하였다[1]. 이러한 문제는 정치적 무관심으로 이어질 수 있다.



(그림1) 정치 참여를 어렵게 하는 이유

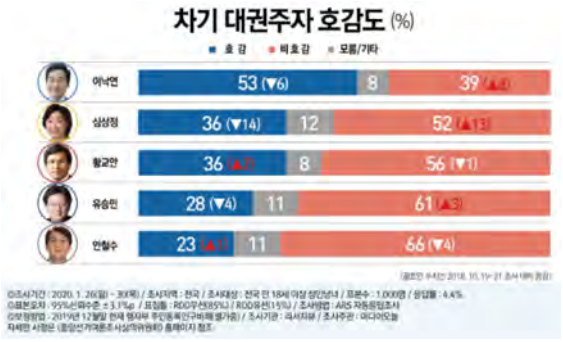
반면에 4차 산업혁명 시대의 도래로 인해 언제 어디서든지 다양하고 흥미로운 정보를 얻을 수 있다. 이러한 점들을 바탕으로 정치 분야도 다소 변화를 겪게 되었다. 선거 연령이 낮아짐에 따라 스마트 기기나, SNS를 즐기는 층에서는 다소 흥미가 떨어지고 어려운 정치보다는 연예 기사와 같은 쉽고 재밌는 기사에 관심이 더 쏠리게 된다. 또한 SNS에 떠돌아다니는 불확실한 정치 관련 정보를 통해 편향된 정치 성향을 갖는 위험을 초래할 수 있다.

본 연구에서는 특정 정치인의 뉴스 기사와 댓글인 빅데이터를 수집하였다. 기사의 카테고리별 분류, 핵심문장 추출을 통해 시간 흐름에 따른 정치인의 견해 변화와 흐름을 쉽게 파악할 수 있게 타임라인을 작성하고 댓글 긍정, 부정 의견 분석을 진행하여 정치인의 평판을 한눈에 볼 수 있도록 하였다. 이를 통해 정치 이슈에 대한 접근성을 향상시켜 올바른 선거권을 행할 수 있게 설계하였다.

이를 위해 네이버 기사 2016년 1월부터 2020년 1월까지 약 4년 동안의 데이터를 수집했으며[2], 댓글의 경우 다른 성향의 매체로 네이버와 다음을 선정하여 2019년 9월부터 2020년 2월까지의 댓글을 수집하여 데이터를 분석하였다[3].

2. 연구 구성

본 연구에서는 여론조사 전문기관 '리서치뷰'에서 시행한 호감도 조사 결과에 따라, 적용 샘플로 이낙연 정치인을 프로파일 대상으로 선정하였다.



(그림2) 차기 대권주자 호감도 여론조사

아래는 본 연구의 결과물로 대상의 간략 소개 및 기사 분석을 통해 파악할 수 있는 정치 행적을 대선, 통일, 외교, 노동이라는 4가지 주제의 타임라인 형태로 나타내고 있으며 댓글 분석을 통한 여론 평판을 보여주고 있다.

(그림3) 본 연구의 결과물

2.1 기사 수집과 데이터 전처리

이낙연 정치인의 행동/의견 타임라인을 제작하기 위해 기사를 수집하였다. 특정 언론사의 정치 성향에 영향을 받지 않기 위해 여러 언론사가 모여 있는 네이버 뉴스에서 기사 수집을 진행하였다. 기사 수집 기간은 2016년 01월부터 2020년 1월로, '이낙연' 키워드가 들어있는 기사를 수집해 약 4만 5천 개의 기사를 타임라인 기사 후보로 두었다.

여기서 중요한 점은 올바른 데이터 분석 결과를 얻기 위해서는 올바른 데이터를 입력해야 한다는 것이다. 우수한 분석 알고리즘을 설계하는 것만큼이나 가다듬어진 데이터를 확보하는 것은 중요하다. 이를 확보하기 위해서 충분한 데이터 전처리 과정은 필수적이다. konlpy 라이브러리를 이용하여 한글 형태소 분석을 진행하였다. 그중 Okt(Open Korean Text) 분석기를 유사도 처리 및 댓글 긍 부정에 사용하였다. 이때 불용어 리스트를 만들어 전치사, 관사, 너무 많이 등장하는 단어 등 문장이나 문서의 특징을 표현하는 데 있어서 불필요한 단어를 삭제하는 불용어 삭제 단계도 포함하여 진행하였다. 또한 gensim 라이브러리를 이용하여 Topic Modeling과 Word Embedding 기능을 통해 기사 학습 및 분류 그리고 핵심 문장 추출을 진행하였다. 이렇게 데이터 전처리 과정을 수행함으로써 데이터 마이닝이 정확성 면에서 높은 확률로 분석 결과가 도출되게 연구하였다.

2.2 기사별 유사도 처리

앞에서 수집한 타임라인 기사 후보들 중에서 같은 내용의 기사이거나 유사한 의미를 갖는 중복 기사들을 걸러내기 위해 TF-IDF를 사용하였다[4]. TF-IDF는 텍스트 마이닝에서 문서에서 특정 단어의 빈도수를 이용해 중요도를 나타내주는 값이다. TF는 문서 안에서의 용어의 빈도, IDF는 한 용어가 문서 집합 전체에서 얼마나 공통적으로 나타나는지 나타내는 값으로 TF와 IDF를 곱한 수치를 문서 간 유사도를 파악하는데 사용하였다. TF-IDF를 이용해 80% 이상의 유사도를 갖는 기사들은 삭제하였다. 예를 들어, "이낙연 국무총리와 부인 김숙희 여사가 서울 서대문구 신촌세브란스 병원 장례식장을 찾아 조문하고 있다."와 "이낙연 국무총리와 부인 김숙희 여사가 서대문구 신촌세브란스 병원 장례식장을 찾아 조문한 뒤 유족을 위로하고 있다"의 경우, 앞의 내용은 같지만, 뒤의 기사에서는 앞의 기사 내용에 "유족을 위로하고 있다"라는 내용만 추가되었기 때문에 문서 간 유사도가 높다고 판단되어 삭제된다. 이러한 과정을 통해 약 4만 5천 개의 기사에서 약 2만 5천 개의 기사로 후보를 축소하였다.

2.3 기사 카테고리 분류와 핵심기사 추출

본 논문에서는 기계학습 모델 중 '양방향 LSTM

순환 신경망'을 이용하여 기사를 대선, 통일, 노동, 외교 총 4가지의 카테고리로 분류하였다. 양방향 순환 신경망은 순방향과 역방향의 두 개의 분리된 순환 신경망을 통해 학습을 시키는 방법이다[5]. 역방향인 은닉층에서는 과거의 정보를 기억하여 학습시키므로 언어 간 의미를 파악하기 위한 본 실험에 적합한 모델이다. 기사를 분류하기 위한 트레이닝 셋을 만들기 위해 카테고리별 4천 개, 총 1만 6천 개의 기사를 수집하였다. Tensorflow v.3.7에서 Learning rate는 0.001로 설정하고 5번의 반복 학습을 수행하였다. 그 결과 87%의 정확도를 얻었다. Word2Vec를 이용하여 단어를 수치화하여 단어 간 유사도를 구하고 embedding 파일과 model을 생성하였다.

이후, 카테고리를 분류하는 방법은 분류하고자 하는 기사를 Word2Vec로 토큰 처리하여 생성된 embedding 파일과 Convert2Vec 하였다. 트레이닝 셋을 만들 때와 같은 환경으로 실험하여 96%의 정확도를 얻었다. 위의 과정을 통해 카테고리별로 타임라인 후보 기사들을 분류하였다. 해당 월중에 핵심 기사를 파악하기 위해 TextRank를 사용하였다. TextRank는 워드 그래프나 문장 그래프를 구축한 뒤, 그래프 랭킹 알고리즘인 PageRank를 이용하여 각각 키워드와 핵심 문장을 선택하는 방법이다[6]. 이러한 과정을 통해 각 카테고리별로 타임라인이 완성된다.

(그림4) 타임라인 작성 순서도

2.4. 댓글 학습과 분류를 통한 여론 분석

댓글 상에 나타난 '이낙연'에 대한 긍정과 부정의 여론을 보여주기 위해 2019년 9월부터 2020년 2월까지 6개월간의 '이낙연' 관련 기사의 댓글들을 네이버와 다음에서 각 포털별로 한 달에 5천 개씩을 수집하였다. 수집한 댓글들을 긍정과 부정으로 분류하기 위한 트레이닝 셋을 만들기 위해 정치 분야의 기사 댓글 약 3천 개를 추가로 수집하여 긍정과 부정, 중립으로 분류하였다. '이낙연'에 관한 여론만을 파악하기 위해 이낙연에 대한 긍정 변수 그룹과 부정 변수 그룹을 설정하고, 분류된 댓글 중 긍정 댓글과 부정 댓글의 주어 부분을 변수 '$'로 바꾼 후 변수 그룹을 대입하여 트레이닝 셋을 만들었다.

(그림5) 긍정 변수 그룹과 부정 변수 그룹

(그림6) 긍·부정 파라미터 공식

이렇게 만든 트레이닝 셋을 바탕으로 앞서 기사의 분류에 사용했던 양방향 LSTM 기계학습모델을 통해 embedding 파일과 model을 생성한 후 각 월별 1만 개의 댓글을 분류하였다. 나온 결과에 대해서는 긍정과 부정 각각의 워드 클라우드를 생성하여 댓글에 많이 등장한 키워드를 보여주고, 긍정 대 부정의 비율로 계산하여 그래프로 보여준다.

3. 기대효과 및 향후 계획

본 연구의 결과물은 특정 정치인의 행보를 시간의 흐름에 따라 한눈에 보여주며, 댓글 상에 나타난 특정인에 대한 대중의 긍정, 부정의 의견을 보여준다. 이러한 결과물을 통해 기대할 수 있는 효과는 다음과 같다.

첫째, 유권자들이 올바른 선거권을 행사할 수 있도록 도와준다. 둘째, 정치 관련 이슈에 대한 접근성을 향상시킨다. 셋째, 정치인이 자기 자신의 평판을 점검하는 데에 사용할 수 있다.

향후 이 프로젝트에 대한 과정의 전반을 연결하고 자동화하는 것을 목표로 한다. 현재는 데이터의 수집과 유사도 기반 중복 제거, 분류 등 데이터를 처리하는 과정들이 분리되어있다. 이러한 과정들을 연결하고 자동화하여 데이터의 수집과 처리 사이의 격차를 좁히고, 최종적으로는 사용자들이 원하는 인물, 기간 등을 설정하여 만들어지는 사용자 기반의 프로파일링 프로그램을 구축하고자 한다.

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음(2016-0-00022)

참고문헌

[1] 한국청소년정책연구원, 「청년사회·경제실태조사」

[2] https://news.naver.com/

[3] https://news.daum.net/

[4] 오유리, 민재옥, 김용일, 김대중, 박용균, 이봉건. (2017). 워드임베딩 기반 TF-IDF를 이용한 특허 문서의 유사 청구항 도출 기법 비교 분석. 한국정보과학회 학술발표논문집, 1002

[5] 주일택, 최승호. (2018). 양방향 LSTM 순환신경망 기반 주가예측모델. 한국정보전자통신기술학회 논문지, 11(2), 204-208.

[6] 배원식, 차정원. (2010). TextRank 알고리즘을 이용한 문서 범주화. 정보과학회논문지. 컴퓨팅의 실제 (한국정보과학회), 16(1), 110-114.