선거방송을 위한 선거후보 당선자 예측 어플리케이션 – 제 20 대 국회의원 선거에 적용한 연구 –

양근석, 구진원, 노민철, 신용우
MBC 기술연구소
{gsyang, kuchi, rohmc, ywshin}@mbc.co.kr

Application for Predicting Candidate on Election Broadcasting – A Case Study on the 20th Assembly Election –

Geunseok Yang, Jinwon Gu, Minchul Roh, Yongwoo Shin
MBC Engineering Research Center

요 약

민주주의의 꽃, 제 20 대 국회의원 선거가 막을 내렸다. 지난 선거에서는 방송사뿐만 아니라 정당들도 엄청난 비용 지출과 노력이 소요되었다. 한 예로, 지난 4. 13 총선거 (제 20 대 국회의원)에서 방송 3 사 출구조사 비용으로 약 66 억원 이상이 지출됐다. 그리고 정당에서는 여론조사 비용으로 약 70 억원 이상을 지출했다. 이러한 큰 비용 지출과, 담당자들의 노력을 줄이기 위해 본 논문에서는 텍스트 마이닝과 감정분석을 적용한 후보 당선자 예측 어플리케이션을 제안한다. 첫째, 소셜 그래프 모델을 소개하여 지역 구조를 발견한다. 둘째, 텍스트 마이닝 기법을 이용하여, 후보자 관련 데이터를 가공한다. 셋째, 텍스트 감정 분석을 통해 후보자의 정보를 수치화 한다. 본 논문의 성능과 효율성을 평가하기 위해, 제 20 대 국회의원 선거에 사례연구를 진행하였다. 제안한 방법이 정확도와 수학적 통계 검증을 통해 가치 있는 효율성을 보였다. 선거방송을 위한 후보자 예측 도구의 도입으로 향후 선거(방송)에서의 큰 비용과 노력을 줄이는데 도움을 줄 것이라 기대한다.

1. 서론

최근 방송 IT 기술의 빠른 성장으로, 방송 시스템과 방송 미디어 산업의 규모가 계속 커지고 있다[1]. 국내외 많은 방송 기획자들은 “차별화된 방송 콘텐츠 기획”에 집중하고, 현업 업무 시간의 상당량을 기획에 보내고 있다. 한 예로, 제 20 대 국회의원 선거에서 지상파 방송 3 사(KBS, MBC, SBS)에서 차별화된 선거방송을 위해 출구조사 비용으로 약 66 억원 (조사원 약 1 만 2500 명, 감독관 약 500 명)을 지출하였다[2]. 그리고 정당들도 여론조사를 실시하여 후보 당선자 사전예측을 하고, “차별화된 득표 전략” 준비에 집중한다. 지난 선거에서 정당 여론조사 비용으로만 약 70 억원 이상을 지출했다 (2016 년 3 월 14 일 기준, 새누리당, 더불어민주당) [3]. 만약 자동화된 후보자 예측 및 가이드라인을 제공하는 소프트웨어(도구)가 존재한다면, 담당자들의 업무량과 선거 예산을 절약하고, 독창적인 선거(방송) 기획이 가능하다.

이러한 문제점을 해결하기 위해, 본 논문에서는 텍스트 마이닝 기법[4]과 감정 분석[5]을 이용하여 선거방송을 위한 후보 당선자 예측 도구를 제안한다. 첫째, 소셜 그래프[6]를 소개하여, 과거 당선 정당과 지역 구조(지역-정당)를 발견한다. 둘째, 텍스트 마이닝 기법을 이용하여, 관련 후보자들의 소셜 데이터를 수집 및 가공한다. 셋째, 텍스트 감정 분석을 통해 후보자들의 정보를 수치화 한다. 제안한 선거방송을 위한 예측 도구의 성능과 효율성을 평가하기 위해, 제 20 대 국회의원 선거 사례 연구를 진행하였고, 제안한 후보자 예측 도구가 효율적이라는 것을 보인다.

본 논문의 구성은 다음과 같다. 2 절에서 관련 연구를 살펴본다. 3 절에서 제안한 후보 당선자 예측 도구를 소개하며, 4 절에서 모의 실험을 진행한다. 5 절에서 실험의 결과를 논의하고, 마지막으로 6 절에서 본 논문에 대한 결론을 맺는다.

2. 관련 연구

“차별화된 방송 콘텐츠”를 위해 지상파 방송 3 사에서는 출구조사를 진행한다. 기존 출구조사에서는 개표시 선거구별 개표율의 미반영으로 최종 예측에 혼선이 일어날 수 있어, 베이지안[7] 확률 기법을 이용하여 출구조사 자료와 개표결과를 통합하는 기법을 제안했다[7]. Park[8]은 출구조사와 유권자 표본크기와의 통계적인 분석을 하였다.

일반적으로 소셜 네트워크[6]는 사용자간 행위 연관성을 분석하여, 사용자 맞춤형 콘텐츠 추천 연구로 응용이 가능하다. Lam [6]은 소셜 네트워크 기반 SNACK 이라는 방법을 소개하며 협업 필터링 (CF, Collaborative Filtering) 알고리즘을 향상시켰다.

정보 검색 (IR, Information Retrieval) 기법은 비구조적인 문서에서 정보를 검색한다 [4]. 따라서 문서를 분석할 수 있고 의미 있는 텍스트 분석이 가능하다. Rao [9]는 온라인 뉴스의 텍스트 의미분석을 통하여 감정 사전을 만들었다.

본 논문에서는 소셜 그래프를 통하여 지역-정당 연관성을 분석 후, 후보 당선자 예측에 활용 하고, 감정 사전을 구축/이용하여, 후보자들의 정보를 수치화 한다.

3. 선거방송을 위한 후보자 예측 어플리케이션

이 장에서는 제안한 어플리케이션의 전반적인 세부 흐름을 아래 그림 1 과 같이 소개한다.

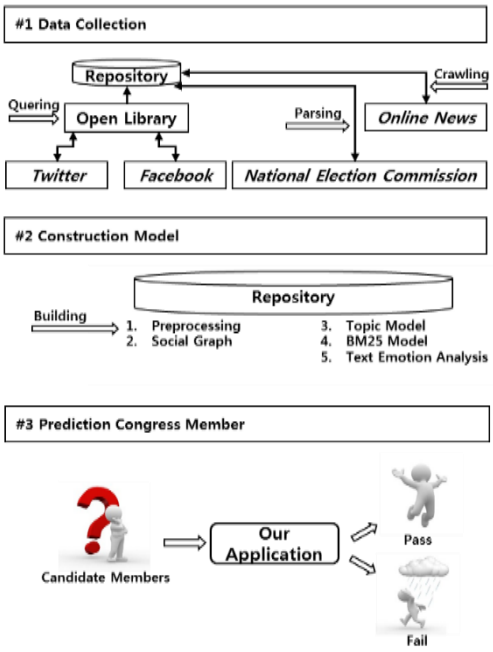

그림 1. 선거방송을 위한 예측 도구의 개요

1) Data Collection

관련 데이터를 수집하기 위해, 통계분석 프로그램으로 자주 사용되는 R 언어[10]를 사용한다. Open Library (TwitteR[11], RFacebook[12])를 이용하여 한국어 공식 홈페이지[13]에서 Query(후보자 키워드)에 대한 사용자 ID, 내용 등을 수집한다. 그리고 온라인에서의 뉴스 콘텐츠를 웹 크롤링[14] (Web Crawling)을 통해 무작위로 데이터(후보자 키워드)를 수집하고, 한국어 형태소[15] 분석을 진행 후 저장소에 보관한다. 마지막으로 과거 선거 통계자료를 수집하기 위해, 중앙선거관리위원회 선거통계시스템[16]에서 데이터를 수집한다.

2) Construction Model

(1) Preprocessing

문서(수집된 데이터)에서 정보를 검색하기 위해 가장 먼저 자연어 전처리 (Natural Language Preprocessing)[17] 과정을 진행한다. 전처리 과정에는 일반적으로 사용하는 정지단어 제거, 원형 추출, 특수문자 제거, 단어 토큰화를 포함한다. 예를 들어, "*# 근석이는 회사에서 밥만 먹는다. ^^*" 라는 문장이 있으면 전처리 후 "근석", "회사", "밥", "먹다"의 단어를 추출한다.

(2) Social Graph

과거 통계 자료를 이용하기 위해 정당과 지역(지역-정당)의 연관성을 저장소에서 추출하고, 아래 식 1 을 진행한다.

$$\text{SGScore}(\text{지역명}, \text{정당명}) = \sum_{n=17}^{19}(Win\_Score) \quad (1)$$

- Win_Score 은 누적 카운트로써 지역에 특정 정당이 당선되었음을 나타낸다.

제 20 대 정당명을 기준으로 과거 정당 변천사를 따른다. 제 1 대부터 제 11 대까지는 선거구와 지역간 예측이 거의 불가능 (예: 제 1 선거구, 제 2 선거구 등)하고, 제 12 대부터 제 16 대까지는 선거구 지역명 통합 분리로 연결에 어려움이 따른다. 따라서 제 17 대부터 제 19 대 국회의원 선거 통계를 사용한다.

(3) Topic Model

저장된 문서들은 한 카테고리가 아닌 의미(주제) 별 다양한 카테고리로 분류될 필요가 있다. 본 논문에서는 LDA (Latent Dirichlet Allocation)[18]를 사용하여 토픽 모델링을 진행한다. 분류될 토픽들은 여러 문서에서 동시 발생한(Co-Occur) 단어들의 분포로 주어진다. 일반적으로 LDA 는 4 개의 파라미터를 가지며, 다음과 같은 특성을 가진다. α와 β는 연관 요소로써, 만약 α가 높다면 하나의 문서가 여러 개의 토픽에 포함될 확률이 높아지고, β가 높다면 하나의 토픽에 다양한 단어들이 포함될 확률이 높아진다. N은 분류될 토픽의 수 이고, R은 반복할 횟수이다. 따라서 저장소에 저장된 문서들로 토픽 모델링을 구축한다.

(4) BM25 Model

BM25[17]는 TF-IDF[17] 기반 기법이며 가중치를 조절하여 효과적으로 텍스트 유사도를 구한다. 일반적으로 BM25 에는 2 가지 파라미터가 있다 (파라미터 표현 기호는 다를 수 있음). k_1 은 단어의 발생(tf, term frequency)에 가중치를 주고, b 는 문서의 길이 관련 가중치를 준다. 따라서 위 가중치를 조절하면, "양근석" 이라는 단어가 "3 번씩" 문서 100 개에서와 다른 문서 10 개에서 나왔다고 가정했을 때, 각 문서 (100 개, 10 개)에서의 "양근석" 단어의 중요성을 다르게 볼 수 있다. 본 논문에서는 후보자 이름을 질의(Query) 하고, 저장소로부터 질의에 대해 가장 유사한 문서들을 추출한다.

(5) Text Emotion Analysis

BM25 의 단계에서 찾아진 문서와 가장 유사한 토픽 문서들을 단어의 매칭 빈도를 이용하여 찾고, 텍스트 감정 분석을 진행한다. 본 논문에서는 감정 사전을 이용하여 수식 2 으로 계산한다.

$$EmoScore(\text{지역명}, \text{정당명}, \text{후보자}) = POS\_TERM\_SCORE + NEG\_TERM\_SCORE \quad (2)$$

- POS_TERM_SCORE 와 NEG_TERM_SCORE 는 후보자가 가진 정보들 중 POSITIVE SCORE (양수), NEGATIVE SCORE (음수)를 의미한다.

3) Candidate Prediction

최종적으로 과거 지역별 당선 통계 (식 1)와 후보자 텍스트 감정 분석 (수식 2)을 결합하여 아래 수식 3 을 계산한다. γ 는 가중치이다.

$$\text{Score}(\text{지역명}, \text{정당명}, \text{후보자}) = \gamma * \text{SGScore}(\text{지역명}, \text{정당명}) + (1-\gamma) * EmoScore(\text{지역명}, \text{정당명}, \text{후보자}) \quad (3)$$

따라서, 선거구별 가장 높은 점수를 가진 후보자가 당선 가능성이 높다고 판단한다.

4. 사례 연구: 제 20 대 국회의원 선거

1) 실험 구성

저장소로부터 지역-정당의 연관성을 소셜 그래프로 찾아내고 식 1 을 진행한다. 저장소 내 문장에 대해 전처리 과정을 진행하고, LDA 를 통해 토픽 모델링을 구축한다. 사용된 파라미터 값은 $\alpha = 0.01$, $\beta = 0.01$, N=30, R=1500 이다 (파라미터 설명은 3.2.3 Topic Model 참조). 제 20 대 국회의원 후보 명단을 이용하여, 후보자 이름과 연관된 유사 문서를 찾는다 (BM25). 사용된 BM25 파라미터는 $k_1 = 1.78$, b = 0.75 이다. 찾아진 유사 문서와 근접한 토픽을 단어의 매칭빈도를 이용하여 유사토픽을 찾는다. SNS 상의 후보자 정보와 유사 토픽 내 문서들에 대해 텍스트 감정 분석을 수식 2 와 같이 계산한다. 상기 모델 구축 파라미터 세팅이 가장 좋은 성능을 보였다. 본 논문에서 사용된 데이터 정보는 다음 표 1 과 같다.

표 1. 사용된 데이터 정보

	Type	Size	etc
Corpus	Construction Model	262,918	~'16/04/12
Emotion Dictionary		98,453	
후보자 / 선거구 수	Test	944 / 253	제 20 대 국회의원 후보
비례대표 후보자 / 선거구 수		158 / 47	

제안한 도구의 성능과 효율성을 평가하기 위해 일반적으로 사용되는 평가척도[19]를 이용한다.

$$\text{Precision(Res)} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall(Res)} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{F} - \text{Measure(Res)} = 2 * \frac{Precision(\text{Res}) * Recall(\text{Res})}{Precision(\text{Res}) + Recall(\text{Res})} \quad (6)$$

- 양근석 후보자가 당선될 것이다라고 예측했고, 실제로 당선자 일 때는 TP, 아닐 때는 FP, 양근석 후보자가 당선자가 아닐 것이라 예측했지만 실제 당선자일 때, FN 을 의미한다.

본 논문은 다음과 같은 연구 질문(RQ, Research Question)으로 진행한다.

- *RQ1: 제안한 선거방송을 위한 후보자 예측 도구가 얼마나 잘 예측하는가?*
- *RQ2: 성능 비교 평가하기 위해, 과거 통계만 고려했을 때 ($\gamma = 1.0$)와 텍스트 감정 분석을 결합한 방법 중 어느 것이 효율적인가?*

이 질문들은 매우 중요한 질문이다. 본 논문은 선거 후보 예측 도구 관련 초기 논문으로써, 현재는 간단한 수식으로 방법론을 정립하고 향후 정치적 요소(투표율/연령 연관성 등)들로 수식 3 을 더욱 구체화할 예정이다.

2) 실험 결과

전국 평균 예측 성능결과를 아래 그림 2 와 같이 보인다. x 축은 파라미터(γ) 수치를 나타내고, y 축은 전국 예측 조화평균(F-Measure)의 비율이다. 예측 조화평균(F-Measure)이 약 70.48%($\gamma = 0.4$)으로 과거 지역-정당 연관성만 고려했을 때($\gamma = 1.0$)보다 좋은 성능을 보인다.

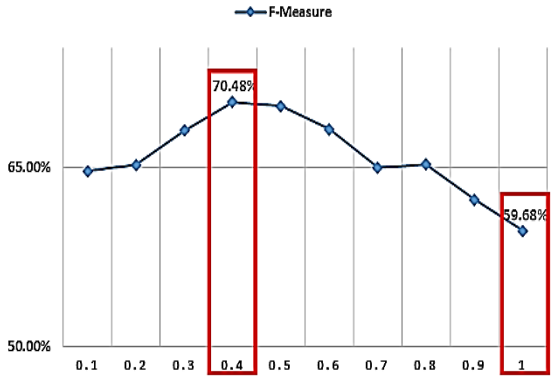

그림 2. 전국 평균 예측 성능 결과 (F-Measure)

AQ1(Answer Question): 각 파라미터 별 전국 예측에서 조화평균 약 70.48%($\gamma = 0.4$)의 성능을 보인다.

RQ2 를 보완하기 위해 수학적 통계 검증[20]을 진행한다. 검증은 R 언어로 진행하였으며, 다음과 같이 가설을 정한다. 귀무가설은 (Null Hypothesis) 제안한 방법이 지역-정당 통계만을 고려했을 때와 통계적으로 유의한 차이가 없음을 보이고 (if P-Value ≥ 0.05), 대립가설 (Alternative Hypothesis)은 두 기법간 통계적으로 유의한 차이가 있음을 보인다 (if P-Value < 0.05). 통계 검증 결과는 다음과 같다.

$H1_0$ (귀무가설)은 Wilcoxon 검증의 P-Value(1.594e-05)을 가지며, 귀무가설을 기각하고, 대립가설을 채택한다 (1.594e-05 < 0.05 is true). 따라서, 제안한 방법이 과거 통계만을 고려했을 때 보다 통계적으로 유의한 차이(신뢰수준 95%)가 있음을 보인다.

AQ2(Answer Question): 제안한 도구가 통계적으로도 유의한 차이(신뢰수준 95%)가 있음을 보인다.

5. 토의

과거 통계만으로 예측했을 때도 전반적으로 양호한 성능을 보였지만, 제 20 대 국회의원 선거 지역 중 호남 지역에 대해 통계만으로는 전혀 예측할 수 없었다 (Recall = 0%). 토픽 모델링 기반 검색 알고리즘과, 텍스트 감정 분석을 결합한 방법이 과거 통계만으로 예측하는 것보다 조금 더 좋은 성능을 보였다. 이러한 결과에 대해 다음과 같이 정밀하게(empirical) 분석을 진행하였다.

먼저, 온라인 상의 문서 (형식이 있는 기사 글 제외)들은 특별한 형식없이 본인의 생각이나 의견을 자유롭게 기술한다. 따라서 온라인 작성자의 지식, 글 표현 능력, 사고방식 등이 콘텐츠 의미 전달에 중요한 역할을 할 수 있다. 따라서 무작위로 파싱된 문서들은 문서 내 내용이 충분할 수 있고, 부족할 수도 있다. 이렇게 무작위로 파싱된 문서들을 토픽 별로 분류하여 의미 있게 형성을 시키고, 그 형성된 문서들끼리는 동질의 특징을 갖게 할 수 있다. 그리고 BM25 의 검색 알고리즘을 통하여 문서 내에 발생한 단어들의 중요도를 다르게 수치화 할 수 있다. 이는 특정 키워드 검색 시 연관성이 높을 수 있는 문서들을 잘 찾을 수 있으며, 후보 예측 성능 향상에 큰 영향을 줄 수 있었다.

본 논문에서는 선거방송을 위한 후보 당선자 예측 도구를 제안하였다. 선거 후보자 예측 관련 초기 연구로써, 사례연구로 제 20 대 국회의원 선거를 진행하였으며, 향후 다양한 요소들을 의미 있게 잘 결합한다면, 성능이 향상 될 수 있는 가능성을 확인하였다 (파라미터 γ). 하지만 과거 통계만 고려 했을 때보다 성능이 향상되었다고 (AQ 1, AQ 2) 다른 선거 (대통령 선거, 국회의원 선거, 지방 선거, 재 선거, 보궐 선거 등) 예측에서도 잘 예측할 수 있다고 일반화 하기는 어렵다 (내부 프로세스, 데이터 등이 다르게 관리된다).

6. 결론

차별화된 방송 콘텐츠 제작은 방송사에서의 핵심역량이다. 지난 제 20 대 국회의원 선거 출구조사와 정당 여론조사에서 엄청난 비용을 지출하였다. 이러한 담당자들의 업무량과 비용을 줄이기 위해 본 논문에서는 텍스트 마이닝과 감정분석을 적용하여 선거방송을 위한 후보 당선자 예측 도구를 제안하였고, 사례연구에서 가치 있는 효율성을 보였다.

향후 다양한 요소들 (투표율-나이, 특정 세력과 등)을 추가하여 예측 정확도를 높일 예정이며, 제안한 선거방송을 위한 예측 어플리케이션을 공개하여, 앞으로 차별화된 선거방송에 기여하고자 한다.

참고 문헌

[1] H. Kim, "방송 콘텐츠 추천검색 기술동향," KBS R&D Book, Vol. 1, No. 2, pp. 5-6, 2011.

[2] 방송 3 사 총선 출구조사…사전투표율 반영이 변수 (2016). http://www.hani.co.kr/arti/society/media/737243.html (accessed May, 25, 2016).

[3] 여론조사 경선 비용… 허리 휘는 예비후보. http://news.chosun.com/site/data/html_dir/2016/03/15/2016031500319.html (accessed May, 25, 2016).

[4] E. Greengrass, "Information Retrieval: A Survey," University of Maryland, Baltimore County, 2000.

[5] Online Sentiment Analysis using R (Lecture Notes), http://statmath.wu.ac.at/courses/SNLP/Presentations/DA-Sentiment.pdf (accessed May, 25, 2016).

[6] C. Lam, "SNACK: Incorporating Social Network Information in Automated Collaborative Filtering," Proceedings of the 5th ACM Conference on Electronic Commerce (EC), pp. 254-255, ACM Press, 2004.

[7] Y. D. Lee and J. Park, "Estimating the Interim Rate of Votes Earned Based on the Exit Poll Results during the Coverage of Ballot Results by Broadcasters," Survey research, Vol. 12, No.1, pp. 141-152, 2011.

[8] J. Park and H. M. Bae, "Statistical Study on Sample Size of Exit Poll in 2010 Local Election," Survey research, Vol. 13, No.3, pp. 89-104, 2012.

[9] Y. Rao, J. Lei, L. Wenyin, Q. Li and M. Chen, "Building Emotional Dictionary for Sentiment Analysis of Online News," World Wide Web, Vol. 17, No. 4, pp. 723-742, 2014.

[10] R Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, 2012.

[11] TwitteR, https://cran.r-project.org/web/packages/twitteR/twitteR.pdf (accessed May, 6, 2016)

[12] RFacebook, https://cran.r-project.org/web/packages/Rfacebook/Rfacebook.pdf (accessed May, 6, 2016)

[13] Twitter, http://www.twitter.com, Facebook, http://www.facebook.com

[14] Web Crawling (Lecture Notes), http://www.cs.ucy.ac.cy/courses/EPL660/lectures/chapter6.pdf (accessed May, 6, 2016)

[15] M. Ko and H. Shin, "Grading System of Movie Review through the Use of An Appraisal Dictionary and Computation of Semantic Segments," Korean Journal of Cognitive Science, Vol. 21, No. 4, pp. 669-696, 2010.

[16] National Election Commission, http://info.nec.go.kr

[17] Y. Tian, D. Lo and C. Sun, "Information Retrieval Based Nearest Neighbor Classification for Fine-Grained Bug Severity Prediction," Proceedings of the 19th Working Conference on Reverse Engineering, pp. 215-224, 2012.

[18] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[19] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," Lecture Notes in Computer Science, Vol. 3408, pp. 345-359, 2005.

[20] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, Vol. 1, No. 6, pp. 80-83, 1945.