

온라인 여론을 수집 및 분석하여 제 7회 전국동시지방선거 서울시장 후보자 득표율 예측 프로젝트

Big Data Business Statistics - Capstone Design

20132605 안상준

Abstract : 각 후보 별 온라인 여론을 조사하여, 패턴을 분석해낸 후, 시계열 데이터 생성 및 변수를 생성하고, 회귀모형을 통해 실제 선거 이후 확인할 수 있는 득표율을 예측하였다.

Keywords : 지방선거, 서울시장, 온라인 여론, 크롤링, 네이버, 트위터, 시계열 데이터, 단일자기회귀, 다변량자기회귀

1. 주제 설정 배경

2018년 제 7회 전국동시지방선거가 예정되어 있다. 4 차산업혁명이 화두인 현재 사회에서, 선거 예측에 빅데이터를 활용할 수 있을까 하는 의문점이 생겼다. 현재 선거 예측은 주로 출구조사 및 여론조사를 바탕으로 진행되고 있지만 여론조사의 한계점이 문제점으로 제기되고 있다. 최근 영국에서 EU 연합 탈퇴를 논의하던 ‘BREXIT’ 사건에 대해 여론조사 결과, EU 연합을 탈퇴하지 않을 것이라 예측하였지만, 실제 개표 후 EU 연합을 탈퇴하는 결과가 일어났다. 이는 설문조사 방식이나 표본 설정이 판세를 정확히 반영하지 못했기 때문이라고 분석된다. 또한, 국내 대선에서도 여론조사의 결과는 실제 투표 결과와는 차이가 있으며, 여론조사의 한계점이 드러났다. 최근 미국 대선에서는, 수많은 여론조사 기관들이 힐러리 후보가 당선될 것이라고 예측하였지만, 실제로는 트럼프가 당선이 되었다. 이는 주로 ‘측정기법의 문제’ 혹은 측정하는 대상인 ‘대중’의 행간을 읽지 못하는 경우로 판단된다. 예측에 실패한 여론조사에 비해, 미국의 한 통계학자인 ‘네이트 실버’는 2008년 대선에서 미국의 50개 주 중 49 개 주의 결과를 빅데이터 분석을 통하여 정확히 예측한 사례가 있다. 그는 여론조사 결과를 활용하되, 온라인에서 각 사용자의 흔적들을 파악하고, 이를 정치적 관념으로 연결하여, 실제로 어떠한 후보 및 정당에 대해 투표를 할지 예측하였다. 이 외에도 해외에서는 선거에 관련되어 빅데이터 분석이 활발하게 이루어지고 있지만, 국내에서는 저조하다는 자료를 확인하였다. 따라서, 빅데이터분석을 통하여 선거 결과를 예측해보는 것을 프

로젝트 주제로 선정하게 되었다.

2. 선행 연구

선거와 연관되어 있는 빅데이터 분석을 활용한 논문을 찾아보았다. 이 주제에 대해서 적지 않은 논문을 발견하게 되었지만, 주로 참고한 논문은 ‘A Multifaceted Approach to Social Multimedia-Based Prediction of Elections’, ‘다차원 가우시안 프로세스와 시계열 텍스트 데이터 이용한 대통령 후보자 지지율 분석’이다. 위 두 논문은 모두 데이터를 수집하여 선거 후보자의 지지율을 예측해 낸 공통점이 있으며, 전자는, Flickr 데이터를 이용하였고, 후자는 뉴스기사 데이터를 중점적으로 활용하였다.

‘A Multifaceted Approach to Social Multimedia-Based Prediction of Elections’ 논문은 미국 대선 및 하원선거를 유의하게 예측하기 위해 VAR 모델에서 수정 및 보완을 한 CVAR (Competitive Vector Auto Regression Model) 을 이용하였다. 이들은, Flickr 내 공유된 선거 후보자 사진의 감정을 추론하였고, 관련 이미지에 대해 사용자의 의견을 분석하여 감정을 추론하였다. 이들은 여론조사의 결과와 Flickr 데이터를 이용하여 매우 정확한 예측을 해냈다. 예측한 결과는 다음 표를 보면 자세하게 알 수 있다.

Date	TABLE I PREDICTION BY DIFFERENT MODELS ON DIFFERENT DAYS									
	Polling		AR		Flickr-AR		VAR		CVAR	
	Obama	Romney	Obama	Romney	Obama	Romney	Obama	Romney	Obama	Romney
2012/10/01	0.4979	0.5021	0.5033	0.4967	0.5006	0.4994	0.4984	0.5016	0.4982	0.5018
2012/10/23	0.4931	0.5069	0.4980	0.5020	0.5014	0.4986	0.5019	0.4981	0.4932	0.5068
2012/11/07	0.5142	0.4858	0.5089	0.4911	0.5142	0.4858	0.5039	0.4961	0.5142	0.4858

State	Official		AR		Flickr-AR		VAR		CVAR	
	Obama	Romney	Obama	Romney	Obama	Romney	Obama	Romney	Obama	Romney
CO	0.5079	0.4796	0.5073	0.4927	0.5074	0.4924	0.5039	0.4934	0.5031	0.4937
FL	0.5044	0.4956	0.4936	0.5064	0.4928	0.5072	0.4917	0.5083	0.5018	0.4982
IA	0.5247	0.4713	0.5139	0.4897	0.4732	0.4891	0.5115	0.4876	0.5291	0.4709
NC	0.5091	0.4106	0.5149	0.4845	0.5155	0.4858	0.5066	0.5134	0.5018	0.5112
NV	0.5335	0.4665	0.5144	0.4856	0.5144	0.4856	0.5165	0.4835	0.5362	0.4638
OH	0.5107	0.4927	0.5022	0.4978	0.5019	0.4981	0.5013	0.4987	0.5185	0.4815
VA	0.5157	0.4843	0.5022	0.4978	0.5219	0.4781	0.5285	0.4715	0.5460	0.4540
WI	0.5339	0.4661	0.5218	0.4782	0.5219	0.4781	0.5285	0.4715	0.5460	0.4540

예측한 날짜는 후보자들의 TV토론 다음 날짜로 선정하였고, AR, Flickr-AR, VAR, CVAR 4가지 모형으로 예측한 결과, CVAR 모형이 굉장히 높은 정확도로 지지율을 예측하였다. 과연 그들은 어떻게 높은 정확도 수준으로 예측하였을까?

우선 그들이 Flickr 데이터를 선택한 이유는, 바로 데이터의 품질성이다. Flickr 데이터는 생성, 업로드, 공유, 댓글 등 여러 과정을 통해 데이터가 생성된다. 이는 다른 SNS에 비해 좀 더 노력을 투자해야 되기 때문에 품질이 높다고 판단한 것이다. 예를 들어, 트위터인 경우 자신의 의견을 아무 제약없이 그대로 작성되고 삭제되는 휘발성이 강한 성격을 지닌 데이터라 한다면, Flickr 데이터는 사람들과의 공유에 초점이 잡혀있기 때문에 이용자들이 업로드 할 때 상대적으로 신중하게 생각하고 노력하여 작성한다고 판단할 수 있다. 또한 기존의 VAR모형은, 변수간의 상관관계행렬을 이용해서 데이터 간 관계성을 파악하지만, 데이터 간 관계성을 높은 수준으로 반영하지 못하기 때문에, 이를 보완하기 위해 Principal Variables, Supporting Variables를 생성하여 반영한 CVAR모델을 이용하였다. 높은 수준의 데이터 품질성 및 정교한 모형의 생성 결과 높은 수준의 정확도를 가진 예측을 할 수 있다고 생각하였다.

‘다차원 가우시안 프로세스와 시계열 데이터를 이용한 대통령 후보자 지지율 분석’ 논문은, 뉴스 시계열 텍스트 데이터와 여론조사 지지율 데이터를 이용하여 가우시안 프로세스에 적용하였고 이를 통하여 대통령 후보자의 지지율을 분석 및 예측한 논문이다. 이들은 2017년 1월 1일부터 2017년 2월 7일까지, 총 38일 기간 동안 발생한 후보자들의 여론조사 결과값을 중앙선거여론조사심의위원회로부터 수집하였다. 이 결과 총 17명의 후보자에 대한 지지율을 얻을 수 있었다. 또한 같은 기간 동안 생성된 뉴스 기사 중 후보자에 대한 키워드를 포함한 신문 기사만을 수집하였고, 수집한 기사를 형태소 분석을 통해 명사에 해당하는 단어만을 추출하였다. 추출한 단어들의 TF-IDF 값을 계산하고, 이를 설명변수, 여론조사 결과 값을 종속변수로 선정하였고 가우시안 프로세스 모델을 통해 실제 후보자 지지율을 예측하였다. 예측한 방식은 1) 시간에 따른 3차 Linear Regression, 2) 시간에 따른 Gaussian Process Regression, 3) 시간과 상위 50개 단어의 빈도수를 이용하여 Gaussian Process Regression을 진행하였다. 전체 기간 중 앞의 80%의 데이터로 모델을 학습하였고,

뒤의 20% 기간 동안 주어진 데이터로 지지율을 예측하여 실제값과의 MAE, RMSE를 계산하여 정확도를 비교하였다. 세 가지 분석 결과는 다음 그림과 같으며, 시간과 상위 50개 단어의 빈도수를 이용하여 Gaussian Process Regression을 한 결과가 가장 좋은 성능을 나타냈다.

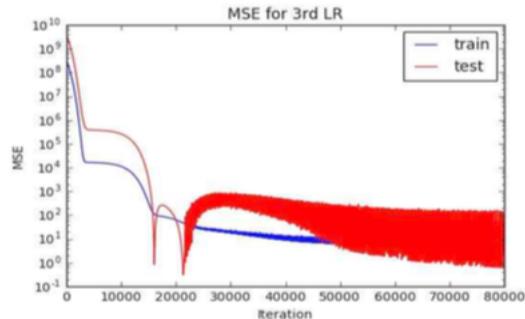


그림 3. 3차 Linear Regression에서 train과 test 데이터의 MSE 변화

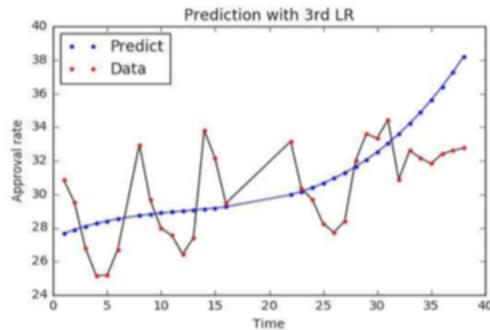


그림 4. 3차 linear regression으로 예측한 결과

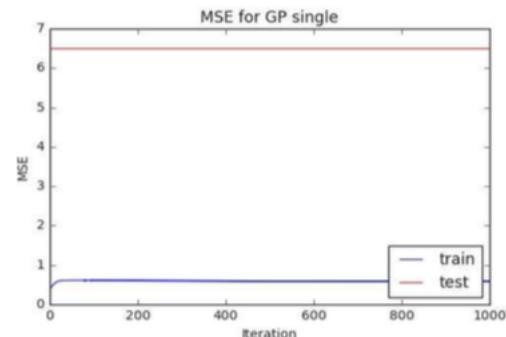


그림 5. 시간에 따른 Gaussian Process에서 train과 test 데이터의 MSE 변화

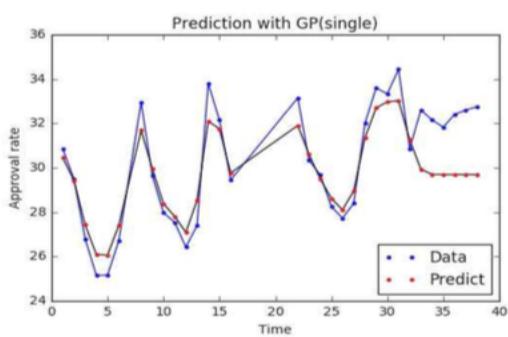


그림 6. 시간에 따른 Gaussian Process에서의 결과

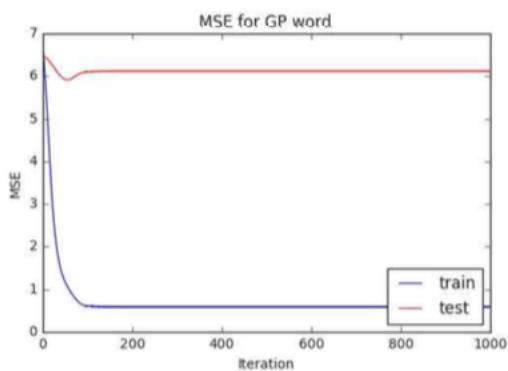


그림 7. 시간과 상위 50개 단어의 빈도수에 따른 Gaussian Process에서 train과 test 데이터의 MSE 변화

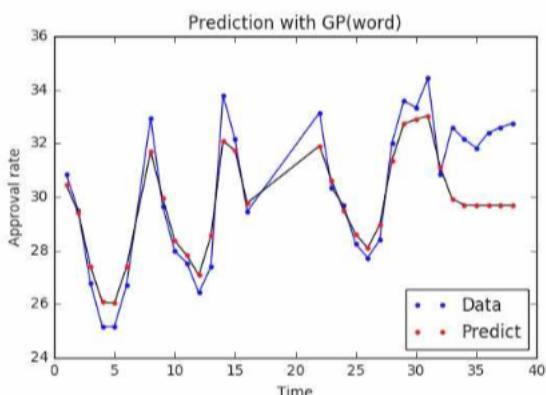


그림 8. 시간과 상위 50개 단어의 빈도수에 따른 Gaussian Process에서의 결과

	MAE	RMSE
3 rd order linear regression	3.93	4.16
GP with Time	2.47	2.54
GP with Time and Words	2.31	2.47

표 2. 세 실험의 MAE, RMSE 표

이 외, 여론조사의 한계점을 극복할 수 있는 방법을 설명한 ‘빅데이터 만능 아니고 여론조사도 언론 활용하기 나름’라는 글을 참고했다. 이 글은, 광운대 미디어학부의 정일권 교수가 작성한 글이다. 이 글에서 강조하는 내용은, 대상자가 온라인 공간에 남긴 이용 흔적을 분석하여 대상자의 생각을 읽어내는 방식으로 진행 시 유의미한 분석이 가능하고, 대상자의 프로파일 정보를 이용하여, 각 이용자의 생각을 파악하는 것이 중요하다는 점이다. 또한, 여론조사의 한계를 빅데이터 분석을 통해 보완할 수 있으며, 선거 결과를 예측할 필요가 있다면, 빅데이터로 전체적인 흐름과, 실시간 변화를 파악하며 원인 분석이나 해석의 근거를 확보하기 위해 여론조사를 활용하는 상호 보완식 이용법을 고려하는 것이 중요하다는 것이다. 위 글의 내용에 따라, 후보자의 선거 결과를 예측할 때 여론조사 결과를 이용하기로 결정하였다.

3. 연구 과정

위의 사례로 든 두 논문 외 연구에서는 보통 뉴스기사, 트위터를 수집하여 분석하였지만, 순수 사용자들의 의견이 반영되지 않았다는 점을 고려하여 댓글을 선정하였다. 앞의 ‘A Multifaceted Approach to Social Multimedia-Based Prediction of Elections’의 논문에서는 데이터의 품질을 중요시하였고, 이와 유사한 우리나라 데이터로는 ‘인스타그램’을 생각하였지만, Flickr 데이터를 활용한 수준까지의 크롤링을 하기에 어려움을 겪어, 데이터의 품질보다는, 순수 사용자들의 의견을 반영할 수 있는 기사 ‘댓글’을 선택하였다. 또한 트위터 역시 휘발성이 강하며 데이터 품질이 상대적으로 낮을 순 있지만, 여론을 순수하게 반영한다는 점을 중요시하여 선정하였다. 정리하자면, 각 후보가 언급된 뉴스 기사의 댓글, 각 후보가 언급된 트위터 데이터를 수집하였다. 지난 2014년 제6회 전국 동시지방선거의 결과값이 존재하므로, 2014년 시기의 데이터를 수집해서 실제 득표율과 어떠한 관계가 있는지 파악하였으며, 이를 통해 2018년 제7회 전국 동시지방선거의 결과를 예측해보았다. 네이버의 뉴스 기사 댓글 경우 2014년과 2018년 모두 크롤링하는데 성공하였지만, 트위터의 경우 오랜 기간이 지난 데이터를 조회하는 것이 불가능하였으며, 2018년에 발생한 데이터만 조회가 가능하였기에, 2014년도는 각 후보가 언급된 뉴스 기사의 댓글을 가지고 진행하였다. 네이버는 N2H4 패키지, 트위터는 Open API를 이용하여 수집하였다.

- 수집 데이터

- 2014년 : 각 후보가 언급된 네이버 뉴스 기사의 댓글
(후보 : 박원순, 정몽준)
- 2018년 : 후보가 언급된 네이버 뉴스 기사의 댓글
+ 후보가 언급된 트위터 데이터
(후보 : 박원순, 김문수, 안철수)

2014년의 경우 4월 1일부터 선거 전날인 6월 3일 까지 수집하였으며, 2018년의 경우 4월 4일부터

선거 전날인 6월 12일까지 수집하였다.

1) 2014년

박원순 후보가 언급된 네이버 뉴스 기사의 댓글
→ 286,533 개

contents	idProvider	regTime	sympathyCount	antipathyCount
OOO들. 국민들이 뭘 원하는지도 모르니?	naver	01/04/2014	3	3
개발이라는 이름으로 전시장들이 낡게 놓은 수많은 건설...	naver	02/04/2014	299	3
오늘의 시장님 어록. "흘러간 물로 물레방아 못 둡...	naver	01/04/2014	81	3
진정한 정치인, 다음 세대를 준비? 정말 오글거린다	naver	02/04/2014	588	45
이제야 본심을 드러내시는군요서민을위한 정책한다고해...	naver	02/04/2014	787	76
철저한 기회주의자	naver	02/04/2014	876	105
정말 공감 된다!!! 김황식 선거본부는 이걸 봐야 할텐데	naver	07/04/2014	2	2
잠실 둘구장 빅점 ☆☆☆☆☆	naver	01/04/2014	2	2
내답시답니다.이상만우절날 여당후보님들의말씀.아...아...	naver	01/04/2014	2	2
AI 칠세가 들와왔구나ㅋ	naver	01/04/2014	685	271

정몽준 후보가 언급된 네이버 뉴스 기사의 댓글
→ 336,700 개

contents	idProvider	regTime	sympathyCount	antipathyCount
OOO들. 국민들이 뭘 원하는지도 모르니?	naver	01/04/2014	3	3
개발이라는 이름으로 전시장들이 낡게 놓은 수많은 건설...	naver	02/04/2014	324	3
오늘의 시장님 어록. "흘러간 물로 물레방아 못 둡...	naver	01/04/2014	87	3
진정한 정치인, 다음 세대를 준비? 정말 오글거린다	naver	02/04/2014	630	45
이제야 본심을 드러내시는군요서민을위한 정책한다고해...	naver	02/04/2014	836	81
철저한 기회주의자	naver	02/04/2014	923	108
새누리당 떡바로해라 저현년문에 새누리 싫어지려한다	naver	01/04/2014	2	2
오사카기부지..서울이....너한..민간인으로,핵을입만개만...	naver	01/04/2014	2	2
1등, 3등 연합으로 2등 정치인문자를 고사시키는 기자....	naver	01/04/2014	3	3
그럼 에이즈 포지티브 해라~	naver	01/04/2014	2	3

2) 2018년

박원순 후보가 언급된 네이버 뉴스 기사의 댓글
→ 406,139 개

contents	idProvider	regTime	sympathyCount	antipathyCount
서울시민들분들은 협영하시죠~	naver	04/04/2018	0	0
사공이 많으니 배가 산으로 가진 않을까....	naver	04/04/2018	1	0
이번엔 무조건 안철수 투표해야 된다. 더불어민주당은 도져...	naver	04/04/2018	0	1
가관이네<U+11A2><U+11A2>ㅉㅉㅉ 신념도 철...	naver	04/04/2018	2	0
병이요	naver	04/04/2018	1	0
연대해도 당선 될 가능성이 회복하니까 안하는 거 다 일지...	naver	04/04/2018	1	0
당마다 깨면서 돌아다니는 걸로도 모자라서 서울시 깨려...	naver	04/04/2018	0	0
안찰스 화이팅 !!!	naver	04/04/2018	0	0
이놈은 도대체 모하는놈인지... 이랬다 저랬다 이번에 떨...	naver	04/04/2018	1	0
안크나이트 찬하다 진짜	naver	04/04/2018	1	1

김문수 후보가 언급된 네이버 뉴스 기사의 댓글
→ 354,960 개

contents	idProvider	regTime	sympathyCount	antipathyCount
서울시민들분들은 협영하시죠~	naver	04/04/2018	0	0
사공이 많으니 배가 산으로 가진 않을까....	naver	04/04/2018	1	0
이번엔 무조건 안철수 투표해야 된다. 더불어민주당은 도져...	naver	04/04/2018	0	1
가관이네<U+11A2><U+11A2>ㅉㅉㅉ 신념도 철...	naver	04/04/2018	2	0
병이요	naver	04/04/2018	1	0
연대해도 당선 될 가능성이 회복하니까 안하는 거 다 일지...	naver	04/04/2018	1	0
당마다 깨면서 돌아다니는 걸로도 모자라서 서울시 깨려...	naver	04/04/2018	0	0
안찰스 화이팅 !!!	naver	04/04/2018	0	0
이놈은 도대체 모하는놈인지... 이랬다 저랬다 이번에 떨...	naver	04/04/2018	1	0
안크나이트 찬하다 진짜	naver	04/04/2018	1	1

안철수 후보가 언급된 네이버 뉴스 기사의 댓글
→ 614,782 개

contents	idProvider	regTime	sympathyCount	antipathyCount
미루기 일찍했군 ☆☆☆	naver	04/04/2018	0	1
안철수 화이팅!!!!!!	naver	06/04/2018	1	0
그이 미지막 나쁜사랑이 성공하길~~~"	naver	06/04/2018	3	0
문술점은 함께 공존할 대상이 아니라는 걸 안철수와 함...	naver	05/04/2018	4	0
김정화 대변인 짓자다 ^^	naver	05/04/2018	7	0
서울시장은 김문수지 철수 따위가 아닙니다.	naver	04/04/2018	0	5
안철수는 다시 할상 나는 안철수!"	naver	04/04/2018	9	1
제발 지지를 10퍼 넘는애들 기사만 써라...뭐 어쩌라는...	naver	04/04/2018	2	5
시장이든 대통령이든 나오는건 지 자유지만...안되는건 안...	naver	04/04/2018	1	3
retire 안철수	naver	04/04/2018	4	2

박원순 후보가 언급된 트위터 개수

→ 373,157 개

x	Date
시장 나가겠다고 한 건 월 일 밤이고 기사가 나온 건 월 ...	01/04/2018
아름다운 양보가 아니라 쇼였다니 안철수 박원순 양보 전...	01/04/2018
시장 나가겠다고 한 건 월 일 밤이고 기사가 나온 건 월 ...	01/04/2018
서울시장 최근 여론조사 결과 박원순 안철수 박영선 안철...	01/04/2018
안철수 박원순 양보 전에 출마결심 접었다 년 월 일	01/04/2018
윤여준 안철수 박원순에게 양보 나흘전에 출마 포기 아버...	01/04/2018
안철수 박원순 양보 전에 출마결심 접었다 년 월 일	01/04/2018
윤여준 안철수 박원순에게 양보 나흘전에 출마 포기 아버...	01/04/2018
구역질 나는 경향신문 기사 민주당 서울시장 예비후보들 ...	01/04/2018
박원순에게 묻는다 박원순에게 묻는다안철수 할부로 발...	01/04/2018

김문수 후보가 언급된 트위터 개수

→ 227,580 개

x	Date
총정육 오세훈 이석연 김병준 김문수까지 자유한국당 흥...	01/04/2018
김문수 반독재 투쟁에 선봉이었다 이제 독재정권의 적자...	01/04/2018
서울 시장에 자한당 후보로 출마할 것으로 보이는 김문수...	01/04/2018
서울 시장에 자한당 후보로 출마할 것으로 보이는 김문수...	01/04/2018
총정육 오세훈 이석연 김병준 김문수까지 자유한국당 흥...	01/04/2018
총정육 오세훈 이석연 김병준 김문수까지 자유한국당 흥...	01/04/2018
총정육 오세훈 이석연 김병준 김문수까지 자유한국당 흥...	01/04/2018
서울시장 후보 김문수 경남도지사 후보 김태호 충남도지...	01/04/2018
김문수 일제 쇠민지가 안됐다면 오늘의 대한민국이 있기 ...	01/04/2018
이번에는 경기도지사도 유능하고 청렴한 사람 뽐내 대...	01/04/2018

안철수 후보가 언급된 트위터 개수

→ 818,085 개

x	Date
안철수 서울시장출마 보도 번째 꼭지 번째 꼭지 번째 꼭...	01/04/2018
시장 나가겠다고 한 건 월 일 밤이고 기사가 나온 건 월 ...	01/04/2018
안희정에서 안철수로 옮기는 전형적인 사례	01/04/2018
아름다운 양보가 아니라 쇼였다니 안철수 박원순 양보 전...	01/04/2018
시장 나가겠다고 한 건 월 일 밤이고 기사가 나온 건 월 ...	01/04/2018
서울시장 최근 여론조사 결과 박원순 안철수 박영선 안철...	01/04/2018
안철수 일 서울시장 출마선언 아름다운 양보 년만에 도전...	01/04/2018
무엇이 보이시나요 안철수를 바라보는 시선에서 국민이 ...	01/04/2018
안철수 박원순 양보 전에 출마결심 접었다 년 월 일	01/04/2018
국민의당 대선후보 안철수의 흑역사 촛불집회와 황제 이...	01/04/2018

또한 2014년, 2018년의 여론조사 결과값을 중앙선거 여론조사 심의위원회에서 수집하였다. 수집된 여론조사 결과값은 2014년 17개, 2018년 29개를 수집하였으며, 겹치는 날짜가 있는 경우, 표본크기를 고려하여 평균치 값을 이용하였다. 수집한 결과는 다음과 같다.

1) 2014년

박원순	정몽준	무응답
20140406	49.3	43.3
20140412	40	34.9
20140430	50.5	37.9
20140505	46.63	35.47
20140512	45.8	30.5
20140513	49.3	35.48
20140514	52.9	32.5
20140517	47.87	34.96
20140519	51	35.4
20140520	46.5	28
20140521	53.5	34.4
20140523	48.58	35.54
20140524	44.3	29.9
20140525	50.6	31.2
20140526	41.44	30.39
20140527	51.58	37.83
20140528	55.73	41.22
		15.65

2) 2018 년

박원순	김문수	안철수	무응답
20180407	50.3	16.6	20.4
20180409	51.5	12.7	21
20180410	54.9	17.1	17.5
20180414	50.9	20.4	19
20180415	51.3	9.5	18.4
20180417	52.1	10.1	13.3
20180503	48.3	9.3	16.5
20180505	42.5	17	21.4
20180506	50.3	10.3	12
20180507	59.5	14.9	13
20180508	56.6	10.6	14.8
20180513	53	10.5	15.2
20180514	52.9	10.9	12.4
20180515	60.8	16	13.3
20180516	56.4	23.5	12.7
20180517	57.3	20.2	12.9
20180518	49	9.9	17.3
20180520	60.1	18.5	12.3
20180521	50.1	11.2	20.2
20180522	51.1	9.1	13.9
20180524	51.2	13.6	15.5
20180527	45.2	15.9	10.9
20180528	46.9	12.9	20.6
20180530	54.2	15.3	13.1
20180531	46.7	27.1	17.4
20180602	56.1	18.2	14.8
20180603	55.5	11.6	14.4
20180604	52.3	13.8	13.7
20180605	44.7	12.3	20.4
			22.6

수집한 데이터를 시계열 데이터로 전처리하여, 네이버인 경우는, 댓글 개수, 공감의 합, 공감의 평균, 공감의 분산, 공감의 표준편차, 비공감의 합, 비공감의 평균, 비공감의 분산, 비공감의 표준편차 변수를 생성하였고, 트위터인 경우 트윗 개수의 합을 변수로 생성하였다. 또한 여론조사 결과값을 활용하기 위해서, 공표되지 않은 날짜는, 전체 데이터를 바탕으로 하나의 다변량 정규분포를 추정하고, 이 분포를 이용해서 다중 대치를 수행하는

“Amelia” 기법을 이용하였다. 전처리가 완료된 데이터는 다음과 같다.

1) 2014 년

- 전처리가 완료된 박원순 데이터

Date	Count	Agree_sum	Agree_mean	Agree_Var	Agree_Sd	Disagree_sum	Disagree_mean	Disagree_Var	Disagree_Sd	research_20145월
2014-04-01	208	87012	418.3269	107942.88	328.5466	35693	171.60396	24303.47	155.8957	46.69349
2014-04-02	430	168590	92.0698	102279.99	319.8124	72122	167.93488	27802.63	166.74712	35.15774
2014-04-03	177	59979	338.8644	91578.40	302.6133	24934	149.87006	25688.47	160.2762	36.08618
2014-04-04	205	93011	311.0001	109432.81	320.2214	80971	104.61379	22358.45	145.5288	40.51924
2014-04-05	774	310663	401.3734	91941.21	303.2214	31557	147.92991	21092.69	161.09884	49.30000
2014-04-06	214	66584	311.1402	110463.63	322.3607	40410	202.33701	33955.63	184.2705	44.80866
2014-04-07	1270	632387	497.9425	104176.38	322.7637	25068	161.30460	28043.41	167.46317	43.44004
2014-04-08	2632	1040676	397.4453	114825.38	338.8189	381802	145.06155	30141.06	173.6118	41.40344
2014-04-09	1671	737133	441.1329	114088.42	327.7712	325930	193.05087	30141.06	173.6118	41.40344
2014-04-10	683	287399	420.7892	109611.72	331.0796	130973	191.76135	29951.66	173.0655	48.64262

- 전처리가 완료된 정몽준 데이터

Date	Count	Agree_sum	Agree_mean	Agree_Var	Agree_Sd	Disagree_sum	Disagree_mean	Disagree_Var	Disagree_Sd	research_20145월
2014-04-01	995	496940	499.4372	104095.33	323.8971	144703	141.43015	25143.09	158.5657	38.35924
2014-04-02	1808	837036	449.3732	115684.62	340.1009	26780	148.10841	21673.31	160.1186	32.95572
2014-04-03	370	157339	425.2405	117386.86	342.6177	59671	161.27297	21619.30	160.0603	28.95192
2014-04-04	210	84710	403.8181	116921.61	310.3105	40410	192.42857	32725.15	180.9000	37.05181
2014-04-05	555	255684	460.6919	104522.00	323.2989	63376	114.19099	24171.54	155.4724	33.27916
2014-04-06	174	50204	388.4748	105615.61	324.9852	8067	161.30460	28099.62	167.6262	43.30000
2014-04-07	181371	506.5497	124235.91	327.7711	86852	239.9265	143.8088	185.2261	141.2249	
2014-04-08	418	172753	341.8274	104176.38	341.2214	88000	121.30432	183.7070	140.8000	36.30400
2014-04-09	188	625577	514.4548	125518.06	344.5915	207981	226.36943	20046.54	176.4595	51.31149
2014-04-10	518	241530	466.2741	127921.96	337.6618	103766	209.32046	26688.54	161.2123	40.32906

2) 2018 년

- 전처리가 완료된 박원순 데이터

Date	Count	Agree_sum	Agree_mean	Agree_Var	Agree_Sd	Disagree_sum	Disagree_mean	Disagree_Var	Disagree_Sd	research_20145월		
2014-04-01	12878	223956	17.837294	14301.44961	181.415202	121667	9.447637	1918.345440	76.930793	4424	18.73440	
2014-04-05	5038	64918	12.849615	1031.62047	48.503629	37728	7.488860	278.475353	16.713835	5277	18.715113	
2014-04-06	239	54817	19.884643	55.352384	30491	40.7040493	402.995827	20.074756	4252	49.15147		
2014-04-07	2461	48428	19.605039	1915.54882	47.769683	20942	8.5095490	151.858140	12.320704	2046	50.30000	
2014-04-08	6113	117909	18.859519	5456.88425	73.870793	57911	9.033697	1124.444861	33.532758	3009	49.17540	
2014-04-09	3107	50012	16.905516	1911.89349	43.725182	32844	10.570988	542.362235	23.288672	2572	51.50000	
2014-04-10	1952	30724	17.859516	1031.62047	48.503629	19471	7.290688	213.412701	14.517704	4487	18.715113	
2014-04-11	9131	183648	16.820544	12308.2785	70.795217	51684	5.9618134	207.331392	7.5115883	3294	50.44143	
2014-04-12	4585	13.202104	11.520508	2851.340302	48.417450	31759	5.1338871	122.664212	8.451773	27.100000		
2014-04-13	1977	81990	12.702040	2026.44412	54.096620	61087	6.7520217	237.167403	48.715508	2309	12.349150	
2014-04-14	763	92212	12.920516	13.959962	2218.03648	318.401885	34242	5.426595	288.479787	16.986993	3802	7.711315
2014-04-15	4973	86793	17.452845	20729.89666	143.978006	27205	5.4705409	1367.418519	36.978595	8290	8.348915	

- 전처리가 완료된 안철수 데이터

Date	Count	Agree_sum	Agree_mean	Agree_Var	Agree_Sd	Disagree_sum	Disagree_mean	Disagree_Var	Disagree_Sd	Twitter_count	support15%
2018-04-04	15901	265186	16.672156	7.93124310	149.7258	148725	9.33134243	4895.367211	69.831980	17272	15.72420
2018-04-05	8079	12.639005	14.62208643	65.511472	47528	158.850324	12.887074	17.727076	15.727706	18.715191	
2018-04-06	4933	91482	18.470327	2715.710529	52.112478	44731	9.031920	38.261957	15.502307	10660	18.718260
2018-04-07	3334	51485	15.458183	1444.248354	38.003268	22529	5.7573485	135.332942	11.632469	5533	20.400000
2018-04-08	7210	132740	18.415045	4974.152975	70.527675	65575	9.0950609	1024.698834	32.019159	6982	18.632027
2018-04-09	4966	83566	16.626621	2773.88112	52.667648	48984	5.9846798	518.517734	22.734734	21.000000	
2018-04-10	19045	19.489009	17.91610412	169.785465	48.417450	151429	5.7951158	128.148891	76.681219	21484	17.500000
2018-04-11	9445	141604	16.839372	5371.665874	73.291649	65645	6.0932182	595.520309	24.403300	18409	20.489236
2018-04-12	7390	131485	16.390263	955.120553	97.74182	59481	8.199359	951.401812	31.486518	13217	18.460887
2018-04-13	11713	20928	17.104947	1958.568640	126.327300	89409	7.633314	277.458511	46.653530	16045	21.750079

전처리가 완료된 데이터에 대해서 ARIMA, VAR 모형을 이용해서 지지율을 예측해 보았다. VAR 모형이란 Vector Auto Regression 의 약자로, 일변량 자기회귀모형을 다변량 자기회귀모형으로 확장시킨 모형이다. 이는 예측 및

을 하고 있어, 구조적 변화가 급속히 진행되어 설명변수의 영향이 변한 경우 이를 적절히 반영하지 못한다는 약점이 있다. 따라서 이러한 시간에 대한 경직성과 주관성을 극복할 수 있는 방법이 ARIMA 모형이라고 할 수 있다. ARIMA 모형은 현재의 관측치 Z_t 는 과거의 어떠한 규칙성에 의해서 재현되며, 이러한 규칙성은 미래에도 유지된다고 가정하고 미래를 예측하고자 했다. 이러한 방법은 모형 설정이 용이한 반면 변수들 사이의 상호작용을 무시하고 있어 일변량 분석이라는 한계에 부딪치게 된다. 이를 회귀모형과 시계열분석의 한계를 보완한 모형이 바로 VAR 모형이다. VAR 모형은 연립방정식 체계와 비슷하나 모형의 오차항을 구조적으로 해석하며 식별제약의 일부가 오차항의 공분산행렬에 가해진다는 특징을 가지고 있어 연립방정식에 비해 유의한 특징을 가지고 있다.

(출처 : 벡터자기모형의 이해 [통계청])

위의 내용을 근거로, 선거 분석 결과를 예측할 때 ARIMA, VAR 모형을 이용하였다.

4. 연구 결과

공표 되어있는 2014년 여론조사 결과값을 이용하여 2014년 실제 득표율을 ARIMA 모형을 통해 예측해 보았다. 단일 시계열 자료이기 때문에, 차분을 통해 안정적인 시계열 데이터로 변환 이후 진행하였다.

1) 박원순

```
> ts_sa <- ts(research_2014$박원순)
> data <- diff(ts_sa)
> auto.arima(data)
Series: data
ARIMA(3,0,0) with zero mean

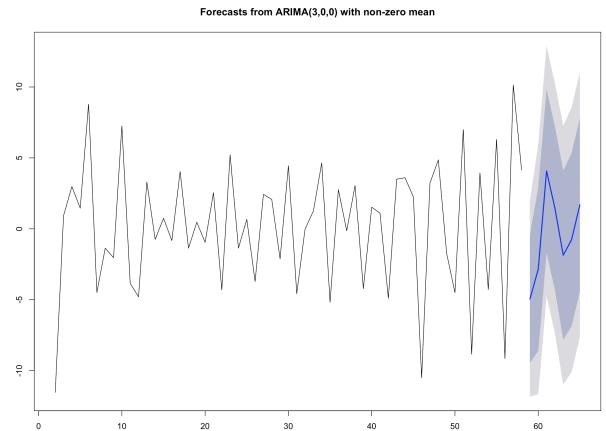
Coefficients:
      ar1      ar2      ar3
    -0.7761   -0.6158   -0.4479
  s.e.  0.1350   0.1607   0.1380

sigma^2 estimated as 13.37: log likelihood=-153.83
AIC=315.66  AICc=316.43  BIC=323.88
> model <- arima(data, order = c(3, 0, 0)) ; model

Call:
arima(x = data, order = c(3, 0, 0))

Coefficients:
      ar1      ar2      ar3  intercept
    -0.7952   -0.6387   -0.4647     0.1951
  s.e.  0.1343   0.1597   0.1368     0.1641

sigma^2 estimated as 12.35: log likelihood = -153.15,  aic = 316.29
> model_forecast <- forecast(model, h = 7) ; model_forecast
Point Forecast    Lo 80     Hi 80    Lo 95     Hi 95
59     -4.953880 -9.458318 -0.4494413 -11.842823  1.935064
60     -2.857984 -8.613030  2.8970619 -11.659568  5.943599
61      4.073396 -1.681720  9.8285122 -4.728295 12.875087
62      1.453721 -4.305483  7.2129248 -7.354221 10.261663
63     -1.863945 -7.817879  4.0899893 -10.969701  7.241812
64     -0.773798 -6.873675  5.3260792 -10.102755  8.555159
65      1.695621 -4.404265  7.7955078 -7.633350 11.024593
```



- ARIMA 모델 결과 예측한 2014년 박원순 후보의 실제 득표율은 52.50313%이다.

2) 정몽준

```
> ts_sa <- ts(research_2014$정몽준)
> data <- diff(ts_sa)
> auto.arima(data)
Series: data
ARIMA(2,0,2) with zero mean

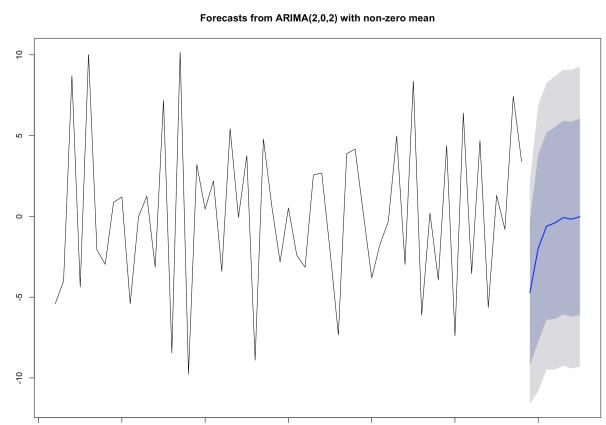
Coefficients:
      ar1      ar2      ma1      ma2
    -0.4729   0.3376   -0.2754   -0.6138
  s.e.  0.2614   0.1632   0.2483   0.2102

sigma^2 estimated as 13.93: log likelihood=-154.71
AIC=319.42  AICc=320.59  BIC=329.63
> model <- arima(data, order = c(2, 0, 2)) ; model

Call:
arima(x = data, order = c(2, 0, 2))

Coefficients:
      ar1      ar2      ma1      ma2  intercept
    -0.4978   0.3198   -0.3308   -0.6692   -0.0683
  s.e.  0.2310   0.1397   0.2208   0.2174   0.0379

sigma^2 estimated as 12.08: log likelihood = -153.73,  aic = 319.45
> model_forecast <- forecast(model, h = 7) ; model_forecast
Point Forecast    Lo 80     Hi 80    Lo 95     Hi 95
59     -4.71828626 -9.210575 -0.2259978 -11.588649  2.152076
60     -1.96663182 -7.752810  3.8195466 -10.815828  6.882564
61     -0.61028149 -6.404022  5.1834595 -9.471044  8.250481
62     -0.40555434 -6.346376  5.5352675 -9.491257  8.680149
63     -0.07373710 -6.061557  5.9140826 -9.231317  9.083843
64     -0.17344826 -6.213529  5.8666324 -9.410954  9.064058
65     -0.01770449 -6.090979  6.0555702 -9.305977  9.270568
> plot(model_forecast)
```



- ARIMA 모델 결과 예측한 2014년 정몽준 후보의 득표율은 33.25436%이다.

실제 득표율은 박원순 후보 56.12%, 정몽준 후보 43.02% 으로, RMSE 값은 7.363745 이다.

실제 득표율과 어느정도 차이가 있기 때문에, VAR 모형을 이용해 보았다. VAR 모형을 이용하기 위해서는 공적분 관계가 존재해야한다. 이를 위해 공적분 검정을 실시하였다.

공적분이란 간단하게 말해서, 시계열자료 내 변수들간의 변동성이 일정수준 이상 벗어나지 않는다는 관계를 의미한다. 공적분 검정을 요한후 검정을 통하여 진행하였다. 박원순 후보와 정몽준 후보의 대표적인 값만 첨부하였다.

- 박원순

```
> mat <- cbind(past_pw[,3], past_pw[,4]) ; mat <- as.data.frame(mat)
> Jo.res <- ca.jo(mat, type = "trace", ecdet = "trend", spec = "longrun") ; summary(Jo.res)

#####
# Johansen-Procedure #
#####

Test type: trace statistic , with linear trend in cointegration

Eigenvalues (lambda):
[1] 4.844808e-01 1.814519e-01 -1.110223e-16

Values of teststatistic and critical values of test:

    test 10pct 5pct 1pct
r <= 1 | 11.21 10.49 12.25 16.26
r = 0 | 48.32 22.76 25.32 30.45

Eigenvectors, normalised to first column:
(These are the cointegration relations)

      V1.l2     V2.l2   trend.l2
V1.l2  1.00    1.00    1.00
V2.l2  12474.89 -85213.79 -49569.63
trend.l2 -85888.96 -26082.23 581174.20

Weights W:
(This is the loading matrix)

      V1.l2     V2.l2   trend.l2
V1.d -1.046954e+00 -7.584073e-02 6.701326e-17
V2.d -1.378728e-05 5.254702e-06 -1.283139e-20
```

- 정몽준

```
> mat <- cbind(past_jm[,7], past_jm[,8]) ; mat <- as.data.frame(mat)
> Jo.res <- ca.jo(mat, type = "trace", ecdet = "trend", spec = "longrun") ; summary(Jo.res)

#####
# Johansen-Procedure #
#####

Test type: trace statistic , with linear trend in cointegration

Eigenvalues (lambda):
[1] 3.064184e-01 2.330863e-01 5.551115e-17

Values of teststatistic and critical values of test:

    test 10pct 5pct 1pct
r <= 1 | 14.86 10.49 12.25 16.26
r = 0 | 35.35 22.76 25.32 30.45

Eigenvectors, normalised to first column:
(These are the cointegration relations)

      V1.l2     V2.l2   trend.l2
V1.l2  1.000    1.00    1.00
V2.l2  4203.811 -121667.59 -16101.06
trend.l2 -43910.596 65136.28 560690.98

Weights W:
(This is the loading matrix)

      V1.l2     V2.l2   trend.l2
V1.d -8.186653e-01 -3.187013e-02 -6.344637e-16
V2.d -9.487763e-06 4.079504e-06 2.742844e-21
```

공적분이 없다고 가정할 때, 검정통계량 값이 유의수준 1% 관점에서 임계값 30.45 보다 크며, 공적분이 존재하는 벡터가 하나 이하로 가정할 시, 검정통계량 값이 유의수준 1% 관점에서 임계값 12.25 보다 작으므로, 각 변수 사이에 공적분 관계가 존재한다고 판단할 수 있다. 따라서 VAR 모형을 이용하기에 적합하다고 판단할 수 있다.

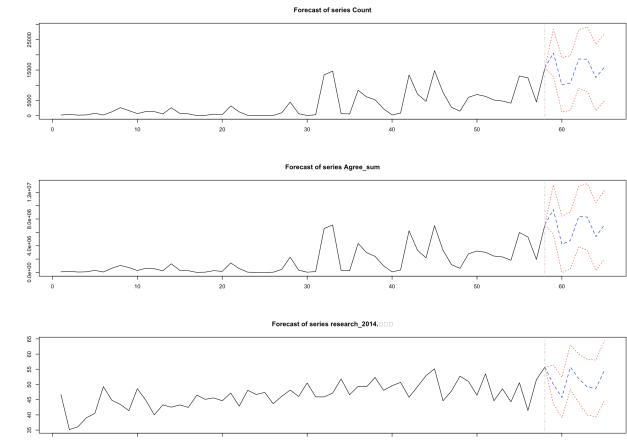
여러 변수의 조합으로 VAR 모형을 진행하였지만, 그 중 댓글개수와 공감지수 합, 지지율의 조합이 가장 높은 정확도를 예측하였다.

- 박원순

```
> res <- VAR(past_pw[, c(2, 3, 11)], 
+             lag.max = 4,
+             ic = "AIC",
+             type = "none")
> prediction <- predict(res, n.ahead = 7, ci = 0.95) ; prediction
$Count
  fcst      lower      upper      CI
[1,] 20624.51 12885.630 28363.40  7738.885
[2,] 10140.06 1177.899 19102.21 8962.157
[3,] 10743.97 1767.146 19720.80 8976.824
[4,] 18641.36 9050.620 28232.10 9590.739
[5,] 18554.63 7870.771 29238.49 10683.857
[6,] 12545.95 1642.656 23449.25 10903.295
[7,] 15906.85 4855.611 26958.09 11051.238

$Agree_sum
  fcst      lower      upper      CI
[1,] 9413388 5733702.44 13093074 3679686
[2,] 4202478 -26862.29 8431819 4229341
[3,] 4782566 546536.49 9018596 4236030
[4,] 8416056 3902316.68 12929795 4513739
[5,] 8285499 3292003.86 13278993 4993495
[6,] 5349432 273056.78 10425808 5076376
[7,] 7098987 1966453.67 12231521 5132534

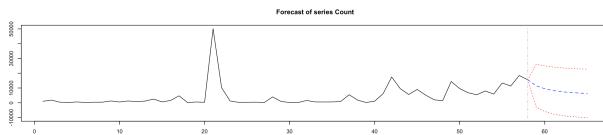
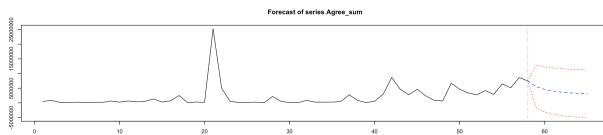
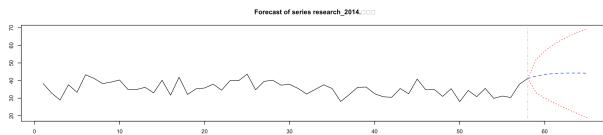
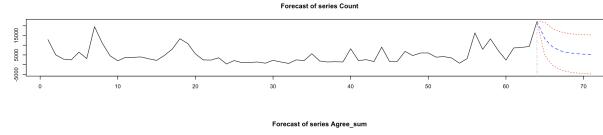
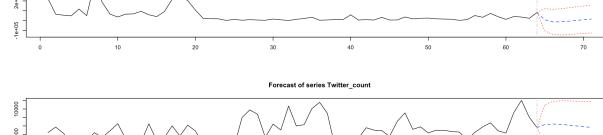
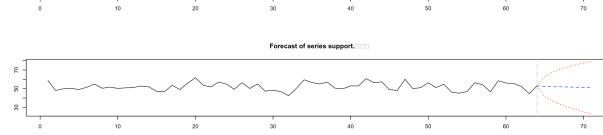
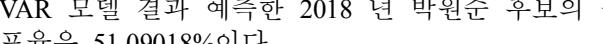
$research_2014.박원순
  fcst      lower      upper      CI
[1,] 49.98606 43.58092 56.39119 6.405135
[2,] 45.74428 39.06082 52.42775 6.683468
[3,] 55.70769 48.37201 63.04337 7.335681
[4,] 51.90691 43.86265 59.95118 8.044266
[5,] 49.11785 39.96811 58.26759 9.149742
[6,] 48.69395 39.33732 58.05058 9.356628
[7,] 54.50316 44.61550 64.39081 9.887651
```



VAR 모델 결과 예측한 2014년 박원순 후보의 득표율은 54.50316%이다.

- 정몽준

```

> res <- VAR(past_jm[, c(2,3,11)],  
+             lag.max = 4,  
+             ic = "SC",  
+             type = "none")  
> prediction <- predict(res, n.ahead = 7, ci = 0.95) ; prediction  
$Count  
    fcst      lower      upper      CI  
[1,] 11479.849 -3068.352 26028.05 14548.20  
[2,] 9511.987 -5961.339 24985.31 15473.33  
[3,] 8283.805 -7603.487 24171.10 15887.29  
[4,] 7480.270 -8627.887 23588.43 16108.16  
[5,] 6948.029 -9294.913 23190.97 16242.94  
[6,] 6591.952 -9745.282 22929.18 16337.23  
[7,] 6350.544 -10060.912 22762.00 16411.46  
  
$Agree_sum  
    fcst      lower      upper      CI  
[1,] 5518708 -1770554 12807971 7289263  
[2,] 4586948 -3121178 12295073 7708126  
[3,] 4007677 -3888299 11903652 7895976  
[4,] 3628754 -4368556 11626064 7997310  
[5,] 3377621 -4682288 11437530 8059909  
[6,] 3209450 -4894755 11313655 8104205  
[7,] 3095280 -5044104 11234665 8139385  
  
$research_2014.정몽준  
    fcst      lower      upper      CI  
[1,] 42.55979 33.02606 52.09351 9.533726  
[2,] 43.43163 30.04822 56.81504 13.383410  
[3,] 43.93249 27.54521 60.31977 16.387281  
[4,] 44.18616 25.22856 63.14376 18.957601  
[5,] 44.27872 23.03617 65.52128 21.242553  
[6,] 44.26671 20.95118 67.58224 23.315531  
[7,] 44.18696 18.96635 69.40757 25.220608  
  
Forecast of series Count  
  
Forecast of series Agree_sum  
  
Forecast of series research_2014.정몽준  
  
  
> res <- VAR(pw[, c(2,3,11,12)],  
+             lag.max = 4,  
+             ic = "AIC",  
+             type = "none")  
> prediction <- predict(res, n.ahead = 7, ci = 0.95) ; prediction  
$Count  
    fcst      lower      upper      CI  
[1,] 13311.340 4794.6434 21828.04 8516.696  
[2,] 9138.758 -343.8302 18621.35 9482.589  
[3,] 7155.473 -2591.7718 16902.72 9747.244  
[4,] 6181.777 -3676.6140 16040.17 9858.391  
[5,] 5678.837 -4250.9890 15608.66 9929.826  
[6,] 5399.495 -4589.3574 15388.35 9988.853  
[7,] 5229.480 -4813.4603 15272.42 10042.940  
  
$Agree_sum  
    fcst      lower      upper      CI  
[1,] 7939.918 -104732.3 120612.1 112672.2  
[2,] -13168.578 -140794.0 114456.9 127625.5  
[3,] -13584.646 -147054.0 119884.7 133469.4  
[4,] -6986.767 -143605.9 129632.4 136619.2  
[5,] 1061.933 -137455.8 139579.7 138517.7  
[6,] 8489.045 -131217.8 148195.9 139706.8  
[7,] 14679.329 -125796.0 155154.6 140475.3  
  
$Twitter_count  
    fcst      lower      upper      CI  
[1,] 6310.423 1614.4351 11006.41 4695.988  
[2,] 6460.723 1141.5157 11779.93 5319.207  
[3,] 6373.097 806.8425 11939.35 5566.254  
[4,] 6196.411 473.7477 11919.08 5722.664  
[5,] 6001.639 153.2445 11850.03 5848.394  
[6,] 5819.152 -140.4081 11778.71 5959.560  
[7,] 5659.381 -402.5405 11721.30 6061.921  
  
$support.박원순  
    fcst      lower      upper      CI  
[1,] 52.48025 41.57367 63.38682 10.90658  
[2,] 52.20524 36.87750 67.53299 15.32775  
[3,] 51.99907 33.31705 70.68109 18.68202  
[4,] 51.79824 30.31497 73.28151 21.48327  
[5,] 51.58171 27.65248 75.51093 23.92923  
[6,] 51.34513 25.22422 77.46604 26.12091  
[7,] 51.09018 22.97189 79.20846 28.11829  
  
Forecast of series Count  
  
Forecast of series Agree_sum  
  
Forecast of series Twitter_count  
  
Forecast of series support.박원순  


```

VAR 모델 결과 예측한 2014년 정몽준 후보의 득표율은 44.18696%이다.

실제 득표율 값은 박원순 후보 56.12%, 정몽준 후보 43.02% 으로, RMSE 값은 1.409959 이다.

2014년 분석 결과, ARIMA 모형보다는 VAR 모형이 좀 더 높은 정확도를 보였으며, VAR 모형내 변수는 댓글 수, 공감지수의 합, 지지율을 선정하였을 때 가장 높은 정확도를 보였다. 이를 바탕으로 2018년에 적용하였다. 2014년도에는 트위터 데이터를 수집하지 못하였으므로, 2018년도에는 트위터 데이터를 추가한 것과 추가하지 않은 것을 비교하였다.

(1) 트위터 데이터를 추가

1) 박원순

VAR 모델 결과 예측한 2018년 박원순 후보의 득표율은 51.09018%이다.

2) 김문수

```

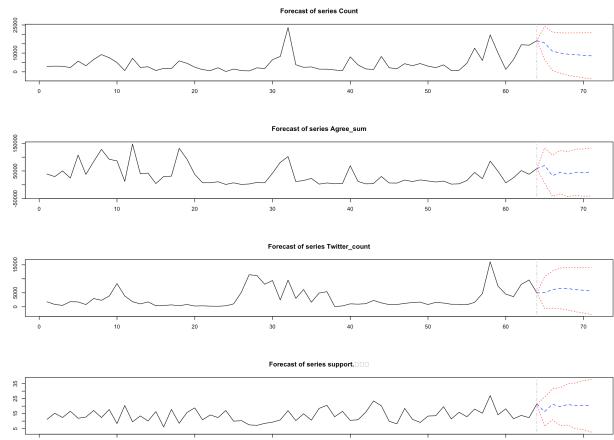
> res <- VAR(km[, c(2,3,11,12)],
+             lag.max = 4,
+             ic = "AIC",
+             type = "none")
> prediction <- predict(res, n.ahead = 7, ci = 0.95) ; prediction
$Count
  fcst    lower    upper     CI
[1,] 15405.070 6507.5970 24302.54 8897.473
[2,] 11009.506  802.0178 21216.99 10207.488
[3,] 10066.099 -709.0041 20841.20 10775.103
[4,] 9409.164 -1885.3647 20703.69 11294.529
[5,] 9125.256 -2541.1467 20791.66 11666.403
[6,] 8817.242 -3184.6291 20819.11 12001.871
[7,] 8550.833 -3735.6738 20837.34 12286.507

$Agree_sum
  fcst    lower    upper     CI
[1,] 69922.23 6362.80 133481.7 63559.43
[2,] 32768.29 -42124.82 107661.4 74893.11
[3,] 45232.50 -33126.20 123591.2 78358.70
[4,] 38967.04 -42921.92 120856.0 81888.96
[5,] 45441.66 -38260.43 129143.8 83702.10
[6,] 43856.98 -41849.58 129563.5 85706.56
[7,] 46324.66 -40757.05 133406.4 87081.71

$Twitter_count
  fcst    lower    upper     CI
[1,] 5048.651 -719.8910 10817.19 5768.542
[2,] 6147.888 -500.0785 12795.85 6647.967
[3,] 6553.328 -724.0542 13830.71 7277.382
[4,] 6493.802 -1138.9229 14126.53 7632.724
[5,] 6112.354 -1810.0366 14034.75 7922.391
[6,] 5894.582 -2249.0930 14038.26 8143.675
[7,] 5631.513 -2709.5193 13972.55 8341.032

$support.김문수
  fcst    lower    upper     CI
[1,] 16.25788 6.549194 25.96656 9.708683
[2,] 21.23095 10.964635 31.49727 10.266319
[3,] 19.68146 6.975507 32.38742 12.705957
[4,] 21.11107 7.345873 34.87627 13.765201
[5,] 20.19164 4.874062 35.50921 15.317573
[6,] 20.67417 4.304229 37.04410 16.369938
[7,] 20.21660 2.668740 37.76446 17.547862

```



VAR 모델 결과 예측한 2018년 김문수 후보의 득표율은 20.21660%이다.

3) 안철수

```

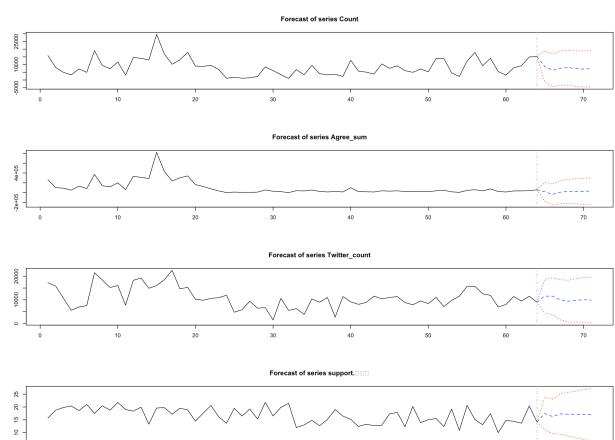
> res <- VAR(ac[, c(2,3,11,12)],
+             lag.max = 4,
+             ic = "AIC",
+             type = "none")
> prediction <- predict(res, n.ahead = 7, ci = 0.95) ; prediction
$Count
  fcst    lower    upper     CI
[1,] 8755.425 -1240.030 18750.88 9995.455
[2,] 6266.843 -4377.079 16910.76 10643.922
[3,] 7672.490 -3417.186 18762.17 11089.676
[4,] 8026.056 -3281.663 19333.77 11307.719
[5,] 7420.577 -4092.508 18933.66 11513.085
[6,] 7115.875 -4551.777 18783.53 11667.652
[7,] 7316.615 -4481.185 19114.42 11797.800

$Agree_sum
  fcst    lower    upper     CI
[1,] 17431.97 -175748.8 210612.7 193180.7
[2,] -33904.20 -250695.2 182886.8 216791.0
[3,] 10023.84 -222129.7 242177.4 232153.5
[4,] 24373.69 -222385.8 271133.2 246759.5
[5,] 26111.53 -229667.0 281890.1 255778.6
[6,] 26638.73 -236574.4 289851.8 263213.1
[7,] 35276.10 -233056.4 303608.6 268332.5

$Twitter_count
  fcst    lower    upper     CI
[1,] 11547.970 4415.9027 18680.04 7132.068
[2,] 11572.409 3803.2559 19341.56 7769.153
[3,] 10025.134 1527.7590 18522.51 8497.375
[4,] 9416.644 530.9268 18302.36 8885.717
[5,] 9819.613 625.8672 19013.36 9193.746
[6,] 10020.946 554.2392 19487.65 9466.707
[7,] 9904.369 185.1626 19623.57 9719.206

$support.안철수
  fcst    lower    upper     CI
[1,] 17.46528 11.212158 23.71840 6.253122
[2,] 16.26933 9.547259 22.99141 6.722075
[3,] 17.29676 9.452407 25.14112 7.844357
[4,] 17.19315 8.743196 25.64311 8.449957
[5,] 17.10731 7.999845 26.21478 9.107468
[6,] 17.05468 7.350922 26.75844 9.703757
[7,] 17.00591 6.751433 27.26039 10.254476

```



VAR 모델 결과 예측한 2018년 안철수 후보의 득표율은 17.00591%이다.

(2) 트위터 데이터를 추가하지 않았을 때,

1) 박원순

```

> res <- VAR(pw[, c(2,3,12)],
+             lag.max = 4,
+             ic = "AIC",
+             type = "none")
> prediction <- predict(res, n.ahead = 7, ci = 0.95) ; prediction
$Count

```

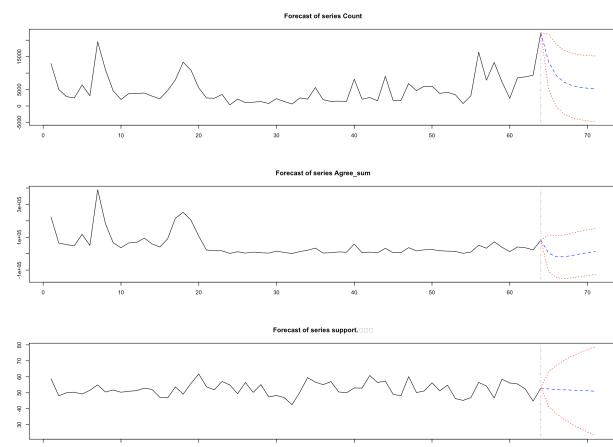
	fcst	lower	upper	CI
[1,]	13412.620	4964.9590	21860.28	8447.661
[2,]	9236.458	-174.3776	18647.29	9410.836
[3,]	7223.649	-2452.2264	16899.52	9675.876
[4,]	6221.101	-3566.0352	16008.24	9787.136
[5,]	5696.658	-4161.6937	15555.01	9858.352
[6,]	5403.110	-4513.8805	15320.10	9916.990
[7,]	5224.417	-4746.1943	15195.03	9970.611

\$Agree_sum

	fcst	lower	upper	CI
[1,]	3245.683	-108846.0	115337.4	112091.7
[2,]	-18902.224	-146226.4	108422.0	127324.2
[3,]	-18750.566	-152013.2	114512.0	133262.6
[4,]	-11053.236	-147437.2	125330.7	136384.0
[5,]	-1877.757	-140086.1	136330.6	138208.3
[6,]	6504.693	-132813.5	145822.8	139318.2
[7,]	13429.655	-126588.9	153448.2	140018.6

\$support.박원순

	fcst	lower	upper	CI
[1,]	52.34036	41.52172	63.15900	10.81864
[2,]	52.02080	36.82380	67.21780	15.19700
[3,]	51.81254	33.29660	70.32849	18.51595
[4,]	51.62571	30.33857	72.91286	21.28715
[5,]	51.42702	27.71987	75.13418	23.70715
[6,]	51.20705	25.33097	77.08313	25.87608
[7,]	50.96581	23.11262	78.81899	27.85319



VAR 모델 결과 예측한 2018년 박원순 후보의 득표율은 50.96581%이다.

2) 김문수

```

> res <- VAR(km[, c(2,3,12)],
+             lag.max = 4,
+             ic = "AIC",
+             type = "none")
> prediction <- predict(res, n.ahead = 7, ci = 0.95) ; prediction
$Count

```

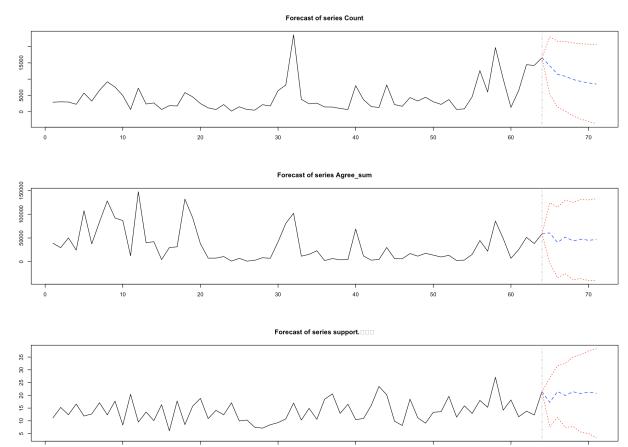
	fcst	lower	upper	CI
[1,]	14122.562	5251.5794	22993.54	8870.982
[2,]	11497.576	1355.9024	21639.25	10141.673
[3,]	10885.466	199.9292	21571.00	10685.536
[4,]	9900.388	-1335.0757	21135.85	11235.463
[5,]	9306.327	-2290.3225	20902.98	11596.650
[6,]	8850.253	-3057.7769	20758.28	11908.030
[7,]	8494.616	-3681.3388	20670.57	12175.955

\$Agree_sum

	fcst	lower	upper	CI
[1,]	60666.05	-3092.049	124424.1	63758.10
[2,]	49355.13	-34251.151	115321.4	74786.28
[3,]	51952.47	-25723.026	129628.0	77675.50
[4,]	43814.53	-37862.863	125491.9	81677.40
[5,]	47539.87	-35823.409	130903.2	83363.28
[6,]	45141.66	-40104.244	130387.6	85245.90
[7,]	46592.94	-39924.766	133110.7	86517.71

\$support.김문수

	fcst	lower	upper	CI
[1,]	17.24411	7.683219	26.80500	9.560891
[2,]	21.51250	11.390900	31.63411	10.121604
[3,]	19.94924	7.386596	32.51189	12.562646
[4,]	21.49178	7.802890	35.00067	13.598888
[5,]	20.74090	5.595775	35.88603	15.145128
[6,]	21.23174	5.012980	37.45050	16.218760
[7,]	20.84601	3.424025	38.26800	17.421985



VAR 모델 결과 예측한 2018년 김문수 후보의 득표율은 20.84601%이다.

3) 안철수

```

> res <- VAR(ac[, c(2,3,12)],  

+             lag.max = 4,  

+             ic = "AIC",  

+             type = "none")  

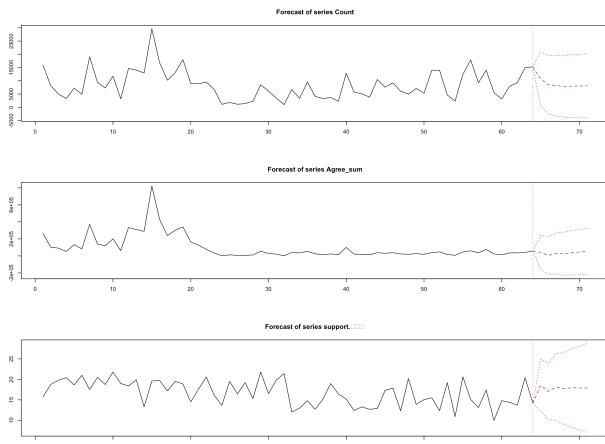
> prediction <- predict(res, n.ahead = 7, ci = 0.95) ; prediction  

$Count
   fcst      lower      upper      CI
[1,] 10722.100 481.0286 20963.17 10241.07
[2,] 8496.859 -2564.1753 19557.89 11061.03
[3,] 8213.387 -3240.2369 19667.01 11453.62
[4,] 7864.746 -3782.4109 19511.90 11647.16
[5,] 7979.841 -3845.3329 19805.02 11825.17
[6,] 8004.434 -3968.4398 19977.31 11972.87
[7,] 8107.276 -4006.5575 20221.11 12113.83

$Agree_sum
   fcst      lower      upper      CI
[1,] 41436.339 -156735.0 239607.7 198171.3
[2,] 7435.913 -214884.9 229756.7 222320.8
[3,] 27653.180 -211917.4 267223.8 239570.6
[4,] 26162.620 -226496.5 278821.7 252659.1
[5,] 39129.430 -221455.2 299714.0 260584.6
[6,] 44994.630 -222041.2 312030.4 267035.8
[7,] 52966.116 -218473.0 324405.2 271439.1

$support.안철수
   fcst      lower      upper      CI
[1,] 18.57367 12.253714 24.89363 6.319958
[2,] 17.05382 10.232030 23.87562 6.821794
[3,] 18.11334 9.979724 26.24695 8.133615
[4,] 17.77408 8.987970 26.56020 8.786114
[5,] 17.95708 8.382226 27.53192 9.574849
[6,] 17.89101 7.664232 28.11778 10.226776
[7,] 17.89020 7.025943 28.75445 10.864253

```



VAR 모델 결과 예측한 2018년 안철수 후보의 득표율은 17.89020%이다.

그렇다면 마지막 여론조사 결과는 어떠할까? 마지막 여론조사 결과는 박원순 후보 52.7%, 김문수 후보 21.6%, 안철수 후보 17.1%로 (주)서던포스트 기관에서 발표하였다. 그렇다면, 실제 득표율과 마지막 여론조사 결과, VAR 모형을 통해 예측한 결과값을 표로 정리해 보았다.

	박원순	김문수	안철수	RMSE
VAR(+T)	51.09018%	20.21660%	17.00591%	2.527195
VAR	50.96581%	20.84601%	17.89020%	2.025647
여론조사	52.7%	21.6%	17.1%	1.746425
실제결과	52.8%	23.3%	19.6%	(6.14 08:00 기준)

각각 RMSE로 평가하였을 때, 여론조사의 정확도가 제일 높았으며, 트위터 데이터를 첨부하지 않은 VAR 모형, 트위터 데이터를 첨부한 VAR 모형 순으로 정확도가 높았다.

5. 한계점

프로젝트를 진행하며 느낀 한계점은, 텍스트 분석을 실시하지 못한 내용이다. 각 댓글별 혹은 트윗별 형태소를 나누거나, TF-IDF, 단어 빈도수의 출현까진 성공하였으나, 감성분석을 실시하지 못하여 댓글별 감성을 부여하지 못하였다. 실제로 각 댓글별로 후보에 대한 긍정 및 부정을 직접 1000개정도 구분하여 CNN을 이용한 Classification을 구성하여 학습시켜봤지만 효과적인 성능이 나타나지 않아 이용하지 못했다. 좀 더 심화적인 텍스트 분석을 진행할 수 있었으면 유의한 예측을 할 수 있지 않았을까 하는 아쉬움이 남는다.

또한, CVAR 논문의 활용이었다. ARIMA, VAR 모형을 이용하여 지지율을 예측하였지만, CVAR 논문의 특징인 Principal Variable, Supportive Variable을 이용하여 변수간의 상관성을 좀 더 정확히 파악하고, 사용자 별 데이터를 이용하여 사용자의 흔적을 분석하여 실제 생각을 유추해내는 단계를 진행하였다면 좀 더 유의한 예측을 할 수 있지 않았을까 하는 아쉬움이 남는다.

참고문헌

- [1] 2006 VAR모형을 이용한 분기GDP 예측모형 연구
- [2] Using flickr for prediction and forecast
- [3] A Multifaceted Approach to Social Multimedia-Based Prediction of Elections
- [4] 다차원 가우시안 프로세스와 시계열 텍스트 데이터를 이용한 대통령 후보자 지지율 분석
- [5] Predicting Election with Twitter: What 140 Characters Reveal about Political Sentiment
- [6] 오피니언 마이닝을 통한 정당지지도 분석 기법
- [7] 빅데이터 만능 아니고 여론조사도 언론 활용하기 나름
- [8] Large Bayesian Vector Auto Regression
- [9] Analyzing and Predicting Youtube Comments and Comment Ratings
- [10] From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series
- [11] 여론조사를 활용한 총선결과 예측모델
- [12] 정치 뉴스 빅 데이터가 선거에 미치는 영향 분석
- [13] 대선후보의 SNS 평판이 선거결과에 미치는 영향 분석 -19 대 대선을 중심으로
- [14] 다수 후보에 대한 선거예측의 정확성 분석
- [15] SNS 감정 분석을 이용한 선거 후보자 순위 예측 시스템
- [16] 텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석
- [17] 20대 국회의원 선거의 정당 지지율 예측을 위한 머신러닝 기법 적용과 문제점 고찰