



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

20대 국회의원 선거의
정당 지지율 예측을 위한 머신러닝 기법
적용과 문제점 고찰

연세대학교 공학대학원
산업정보경영 전공
하 상 현

20대 국회의원 선거의
정당 지지율 예측을 위한 머신러닝 기법
적용과 문제점 고찰



지도교수 김 창 욱

이 논문을 석사학위 논문으로 제출함

2016년 12월 19일

연세대학교 공학대학원
산업정보경영 전공
하 상 현

하상현의 석사학위 논문을 인준함

심사위원 김 창 욱 
심사위원 김 상 욱 
심사위원 김 병 욱 

연세대학교 공학대학원

2016년 12월 19일

감사의 글

반복되는 일상생활 속에서 배움에 대한 열망과 새로운 길에 대한 고민을 하고 있을 무렵 주위의 권유와 고민 끝에 사업과 학업의 병행이라는 새로운 도전을 결심하였고 2014년 9월 대학원 생활을 시작하였습니다. 이렇게 시작한 2년 6개월의 대학원 생활은 저의 인생에 있어서 새로운 활력소가 되었으며 평생 잊지 못할 추억과 좋은 사람들과의 인연을 만들 수 있었던 소중한 시간이었습니다.

먼저 부족한 저에게 많은 가르침과 격려를 아낌없이 주시고, 열정적인 논문 지도를 해 주신 김창욱 교수님께 깊이 감사 드립니다. 아울러 지난 학기 전공대표로서 전공을 이끌어 나갈 수 있는 기회를 주시고 바쁘신 중에도 부족한 저의 논문을 심사해 주신 정봉주 주임교수님과 조영상 교수님께도 깊이 감사 드립니다.

마흔다섯이라는 나이에 다시 시작한 학창 생활을 무사히 마칠 수 있었던 것은 87기 동기들이 함께 하였기에 가능하였습니다. 항상 든든한 힘이 되어 준 성문, 재홍, 세정, 상일, 민철, 병호, 성준, 혁민, 수호,

재일이에게 감사의 마음을 전합니다. 졸업 후에도 평생 좋은 인연으로
이어 가도록 하겠습니다.

언제나 많은 조언과 도움을 주신 창세기랩의 선후배님들과 대학원
생활을 함께 했던 선후배님들께 감사 드리며, 지난 학기 많은 도움을 준
태준, 선영, 동익에게도 감사의 마음을 전합니다.

늦은 나이에 공부한다고 항상 아들 걱정하시는 어머니, 대학원 진학에
많은 조언과 물심양면으로 도움을 주신 누님, 늦은 나이 대학원 진학에도
언제나 믿고 남편을 위해 기도해 준 아내에게 사랑의 마음을 담아 깊이
감사 드립니다.

2016. 12월

하상현 드림

차 례

| | |
|-----------------------------------|-----|
| 그림 차례 | iii |
| 표 차례 | iv |
| 국문 요약 | v |
| | |
| 제 1 장 서론 | 1 |
| 1.1 연구 배경 | 1 |
| 1.2 연구 목적과 의의 | 4 |
| 1.3 연구 방법론 | 5 |
| | |
| 제 2 장 선행 연구 | 7 |
| 2.1 고전적인 감성분석 기법 | 7 |
| 2.2 감성분석을 위한 머신러닝 알고리즘 소개 | 8 |
| 2.3 머신러닝 알고리즘을 이용한 감성분석 사례 | 10 |
| | |
| 제 3 장 연구 방법 | 12 |
| 3.1 데이터 수집 및 전처리 방법 | 12 |
| 3.2 Lasso 및 Ridge 선형회귀 알고리즘 | 15 |

| | |
|-------------------------------|-----------|
| 3.3 검증방법 | 17 |
| 제 4 장 실험 결과 및 분석 | 20 |
| 4.1 20대 국회의원 선거 트위터 데이터 | 20 |
| 4.2 트위터 감성분석 결과 | 21 |
| 4.2.1 새누리당 결과 및 해석 | 22 |
| 4.2.2 더불어민주당 결과 및 해석 | 26 |
| 4.2.3 국민의당 결과 및 해석 | 31 |
| 4.2.4 정당 지지율 예측 결과 및 해석 | 35 |
| 제 5 장 결론 및 한계 | 39 |
| 5.1 연구 결론 | 39 |
| 5.2 연구 한계 | 40 |
| 참고문헌 | 41 |
| [ABSTRACT] | 43 |

그림 차례

| | |
|-----------------------------------|----|
| 그림 1. 연구 방법의 과정 | 6 |
| 그림 2. 서포트 벡터 머신 | 9 |
| 그림 3. 트위터 API를 이용한 데이터 수집 | 12 |
| 그림 4. 트위터 데이터 형태소 분리 | 13 |
| 그림 5. 트위터 데이터의 극성 태깅 | 14 |
| 그림 6. 트위터 데이터의 행렬화 | 15 |
| 그림 7. Lasso 및 Ridge 회귀 수행코드 | 19 |
| 그림 8. 각 정당별 트위터 데이터 분포 | 21 |
| 그림 9. 정당 지지율 비교(Lasso 회귀) | 36 |
| 그림 10. 정당 지지율 비교(Ridge 회귀) | 37 |

표 차 례

| | |
|---|----|
| 표 1. 새누리당 트위터 감성분석 정확도 결과 | 22 |
| 표 2. 새누리당 Lasso 결과 긍정적인 영향을 미친 주요 단어 | 23 |
| 표 3. 새누리당 Lasso 결과 부정적인 영향을 미친 주요 단어 | 24 |
| 표 4. 새누리당 Ridge 결과 긍정적인 영향을 미친 주요 단어 | 25 |
| 표 5. 새누리당 Ridge 결과 부정적인 영향을 미친 주요 단어 | 26 |
| 표 6. 더불어민주당 트위터 감성분석 정확도 결과 | 27 |
| 표 7. 더불어민주당 Lasso 결과 긍정적인 영향을 미친 주요 단어 | 28 |
| 표 8. 더불어민주당 Lasso 결과 부정적인 영향을 미친 주요 단어 | 29 |
| 표 9. 더불어민주당 Ridge 결과 긍정적인 영향을 미친 주요 단어 | 30 |
| 표 10. 더불어민주당 Ridge 결과 부정적인 영향을 미친 주요 단어 | 30 |
| 표 11. 국민의당 트위터 감성분석 정확도 결과 | 31 |
| 표 12. 국민의당 Lasso 결과 긍정적인 영향을 미친 주요 단어 | 32 |
| 표 13. 국민의당 Lasso 결과 부정적인 영향을 미친 주요 단어 | 33 |
| 표 14. 국민의당 Ridge 결과 긍정적인 영향을 미친 주요 단어 | 34 |
| 표 15. 국민의당 Ridge 결과 부정적인 영향을 미친 주요 단어 | 35 |

국 문 요 약

20대 국회의원 선거의 정당 지지율 예측을 위한 머신러닝 기법 적용과 문제점 고찰

연세대학교 공학대학원
산업정보경영 전공
하 상 현

소셜 미디어(Social Media)를 통해 대규모의 데이터가 생성되고 공유되면서 데이터를 효율적으로 저장하고 분석할 수 있는 빅데이터(Big Data) 기술에 대한 관심이 점점 높아지고 있다. 또한 소셜 미디어는 단순한 정보의 공유를 넘어 현실정치와 선거에서 강력한 영향력을 미치는 또 하나의 미디어로서 관심을 받고 있다.

빅데이터 분석 기술의 발달과 소셜 미디어의 영향력이 커짐에 따라 소셜 미디어에서 실시간으로 생성되는 대량의 데이터를 통해 사람들의 감정, 태도, 평가, 의견 등을 분석하는 감성분석(Sentiment Analysis)에 대한 연구가 활발하게 이루어지고 있다.

본 연구에서는 20대 국회의원 선거 관련 트위터(Twitter) 데이터의 머신러닝(Machine Learning) 기법을 이용한 감성분석을 실시하여 정당 지지율을 예측하고 예측모형의 극성분석 정확도를 측정하고자 한다. 또한 예

측모형의 정확도에 문제점이 있다면 문제점을 파악하고 향후 예측 모형의 방향에 대해서 제안하고자 한다.

본 연구를 위하여 사용된 실험 데이터는 20대 국회의원 선거의 공식 선거운동 기간인 2016년 3월 31일부터 2016년 4월 12일까지 "총선", "국회의원", "선거", "413", "새누리당", "더불어민주당", "더민주", "새누리", "국민의당", "공천" 등 10개의 키워드로 총 1,048,574건의 트위터 데이터를 수집하여 구축하였다.

이 연구를 통하여 Lasso 및 Ridge 회귀 모형의 클래스 불균형 문제(Class Imbalance Problem)와 우리나라의 정치문화 여건 상 예측의 부정확성과 트위터의 제한적인 연령대 주된 사용과 정당이나 지지하는 후보를 홍보하는 용도로 사용하는 경우가 많음으로 인한 예측의 부정확성이 나타난 것을 확인하였다. 이러한 예측의 부정확성을 해소하기 위하여 트위터의 특성과 우리나라 정치문화 여건에 맞는 정교한 트위터 감성분석 모델에 대한 연구가 필요하다고 판단된다.

본 연구에서 제안된 방법이 정치 여론의 감성분석 외에 정책 제안이나 회사 제품에 대한 소비자 반응과 같은 감성분석이 요구하는 다양한 다른 영역에서 적용될 수 있기를 기대한다.

주제어: 지지율, 국회의원선거, 소셜 미디어, 트위터, 감성분석, 머신러닝, 극성분석, Lasso 회귀, Ridge 회귀

제 1 장 서 론

본 장에서는 본 논문의 연구 배경을 제시하고 이를 위한 연구의 목적과 방법에 대해 알아보며, 전체적인 논문 구성을 살펴본다.

1.1 연구 배경

소셜 미디어(Social Media)를 통해 대규모의 데이터가 생성되고 공유되면서 대용량의 데이터를 효율적으로 저장하고 분석할 수 있는 빅데이터(Big Data) 기술에 대한 관심이 점점 높아지고 있다.

빅데이터는 통상적으로 데이터의 수집 및 저장, 관리 그리고 분석과 관련된 소프트웨어적으로 수용할 수 있는 한계를 넘어서는 크기의 데이터를 말하며, 빅데이터의 사이즈는 단일 데이터 집합의 크기가 몇 십 테라바이트(Terabyte)에서 수 페타바이트(Petabyte)에 이르며, 그 크기가 끊임없이 변화하는 것이 특징이다(McKinsey, 2011). 2012년 Gartner는 빅데이터를 용량이 크고 속도가 빠르며, 높은 다양성을 가진 데이터라고 재정의 하였다.

최근 빅데이터는 대용량 데이터의 수집과 저장, 그리고 분석, 체계화의 도구, 분석기법과 플랫폼 등을 넓게 포함하는 용어로 변화하고 있으며, 대용량 데이터를 활용, 분석하여 의미 있는 정보를 추출하고 구축된 지식을

기초로 하여 능동적으로 반응하거나 변화를 예측하기 위한 정보화 기술을 말한다(강만모, 김상락, 박상무, 2011).

다양한 대규모 데이터의 생성, 수집, 분석하는 빅데이터 기술의 발전은 현대 사회를 더욱 정확하게 예측하고 효율적으로 작동케 하여 현대 사회의 구성원들에게 맞춤형 정보를 제공하고 관리, 분석이 가능하게 하였다. 빅데이터의 활용이 기업의 ‘빅데이터 경영’을 넘어 대국민 공공 서비스의 영역으로까지 확대되어 가고 있다. 빅데이터가 단순히 기업 경쟁력 향상의 수단에 그치지 않고 국가 경쟁력을 높이는데도 이용되고 있다. 해외에서는 빅데이터와 시스템 통합(System Integration)이 결합된 새로운 공공 서비스 모델이 제시되며 사회 구성원들에게 좋은 품질의 서비스를 제공하고 있다(이성훈, 2012).

최근에는 스마트폰(Smart Phone)의 확산으로 인한 모바일 환경의 발달로 인해 전 세계적으로 소셜 미디어 이용자들이 급증하면서 우리 사회에 일어나는 모든 이슈들에 대한 정보가 대량으로 생산되고 공유되고 있으며, 사회, 문화, 경제와 더불어 정치영역에서도 커다란 영향을 미치고 있다. 트위터(Twitter), 페이스북(Facebook) 등 소셜 미디어에 남긴 정치, 경제, 사회, 문화에 대한 메시지는 그 시대의 감성과 정서를 파악할 수 있는 원천으로 등장하였으며, 개인과 기업은 소셜 미디어를 통해 정보의 생산함과 동시에 전달하는 역할을 하는 시대가 도래하였다(송영조, 2012). 우리나라에서도 여러 분야에서 소셜 미디어의 영향력이 커지고 있지만 특히 정치분야에서의 영향력은 점차 확대되어 가고 있으며, 정치 참여의 도구로도 적

극 활용되어 가고 있다.

소셜 미디어의 정치적 영향에 대한 대표적인 사례는 2008년 미국 대통령 선거에서 버락 오바마(Barack Obama)가 트위터, 페이스북 등 다양한 소셜 미디어를 효과적으로 선거운동에 활용하여 대통령 선거에서 승리하여 미국 최초의 흑인 대통령이 된 사례에서 찾을 수 있다. 그 이후 2010년 영국 총선과 2011년 튀니지의 ‘재스민 혁명’에서도 소셜 미디어는 정치적으로 커다란 영향을 주었다. 우리나라에서도 2014년 서울시장 선거에서 트위터 사용자들이 생산하는 트위터 데이터에 선거 관련 정치적인 이슈가 충분히 반영되었으며, 언론과 여론과의 관계를 분석함에 트위터는 정치적 영향력에 있어서 충분한 가치가 있는 것으로 확인되었다(한범희, 2014).

이와 같이 소셜 미디어는 단순한 정보의 공유를 넘어 현실정치와 선거에서 강력한 영향력을 미치는 또 하나의 미디어로서 관심을 받고 있다. 특히 소셜 미디어의 대표 서비스인 단문 메시지 서비스인 트위터는 정치적 소통의 수단으로써 정치권의 가장 많은 관심을 받고 있다(금혜성, 2011). 소셜 미디어를 통한 정치 참여와 소셜 미디어가 정치에 미치는 영향에 대한 연구는 꾸준히 계속 진행되고 있다.

1.2 연구 목적과 의의

최근 빅데이터는 정치, 경제, 사회, 문화, 의료 등 여러 분야에서 걸쳐 중요성이 부각되어 각 분야에 적극적으로 활용되고 있다. 특히 정치 분야에서 빅데이터 분석이 활용된 대표적인 사례가 2012년 미국 대통령 선거에서 재선에 성공한 버락 오바마 후보 캠프이다. 버락 오바마 캠프는 다양한 형태의 유권자 개개인의 정보를 확보하고 빅데이터의 분석을 이용하여 유권자 개개인의 성향을 파악하고 성향에 맞춰 접근하는 ‘Micro-targeting’ 방식의 선거운동을 전개하였다(차재필, 2012).

빅데이터 분석 기술의 발달과 소셜 미디어의 영향력이 커짐에 따라 소셜 미디어에서 실시간으로 생성되는 대량의 데이터를 통해 사람들의 감정, 태도, 평가, 의견 등을 분석하는 감성분석(Sentiment Analysis)에 대한 연구가 활발하게 이루어지고 있다.

감성분석이란 상품 및 서비스, 기관 및 단체, 사회적 이슈, 사건 등에 관하여 블로그 또는 트위터, 페이스북과 같은 소셜 미디어 등에 남긴 의견을 수집하고 분석하여 사람들의 감정과 태도, 평가 및 의견 등을 분석하는 빅데이터 분석 기술을 말한다(Liu, 2012). 이러한 감성분석은 온라인 쇼핑몰이나, 영화평, 맛집, 호텔 등의 이용자의 좋고 나쁨의 감성을 분석하여 제품 및 서비스 등에 대한 만족도 평가에 활용되고 있다.

감성분석은 다양한 업종의 기업과 사회 여러 분야에서 활용되고 있으며,

이러한 감성분석에 대한 다양한 연구가 진행되고 있다. 정치적 이슈에 대한 소셜 미디어의 대용량 데이터 감성분석을 통하여 실시간 여론의 흐름을 파악할 수 있다(박중호, 2015).

본 연구에서는 20대 국회의원 선거 관련 실시간 트위터 데이터 수집하여 머신 러닝(Machine Learning)을 이용한 데이터 감성분석을 통해 정당 지지율을 예측하고 예측모형의 극성분석 정확도를 측정하고자 한다. 또한 예측모형의 정확도에 문제가 발견되면 문제점을 파악하고 향후 예측 모형의 방향에 대해서 제안하고자 한다.

1.3 연구 방법론

그림 1은 본 연구 방법의 절차를 보여준다. 연구의 방법은 20대 국회의원 선거 관련 실시간 트위터 데이터 중에"총선", "국회의원", "선거", "413", "새누리당", "더불어민주당", "더민주", "새누리", "국민의당", "공천" 등 10개의 키워드로 공식 선거운동 기간인 2016년 3월 31일부터 2016년 4월 12일까지 수집된 총 1,048,574건의 데이터에 대하여 특수기호, 영문, 숫자 및 관사, 전치사, 접속사, 조사와 같은 불용어를 제거하고 트위터 원문을 각각 형태소 단위로 분리하고 단어 하나 하나를 독립변수로 설정한다.

형태소로 분리된 트위터 데이터 개별의견에 극성변수를 추가하는 극성태깅을 실시하고 각 단어를 독립변수, 종속변수로 처리하는 데이터 행렬의

구성한다.

이렇게 추출된 독립변수와 극성변수에 대하여 Lasso 및 Ridge 회귀분석을 통하여 개별 트위터 데이터 감성분석을 진행하여 극성 예측 모형 생성하고 예측 정확도를 분석하고자 한다. 또한 예측 정확도를 기반으로 하여 모형의 정당성 및 한계점 분석하고 문제점이 있는 경우에는 문제점에 대한 고찰을 진행하고자 한다.

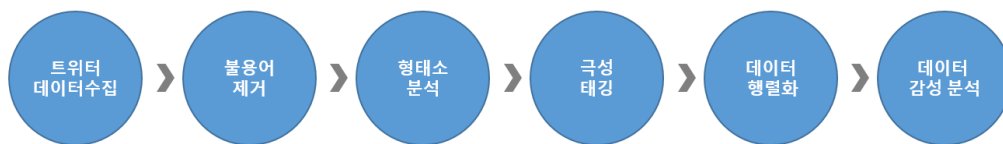


그림 1. 연구 방법의 과정

제 2 장 선행 연구

2.1 고전적인 감성분석 기법

이상훈, 최정, 김종우(2016)는 감성분석에 사용되는 감성사전을 데이터의 특성에 맞게 적절하게 변형하여 구축하는 방법을 시도하였다. 영화 리뷰 데이터를 분석 대상으로 선정하였으며, 대표적 영화정보 사이트 IMDb에서 발생한 약 2년간의 영화리뷰 데이터를 수집·분석하였다. 맞춤형 감성사전 구축을 위한 핵심 기법으로 SO-PMI(Semantic Orientation from Point-wise Mutual Information)를 활용하였으며, 어휘 간 극성이 뚜렷하게 구분되는 형용사에 한정하여 연구를 진행하였으며 분석결과 맞춤형사전을 활용한 감성분석 예측정확도는 영화 장르별로 상이했다. ‘애니메이션’을 제외한 5개 장르에서 기존의 범용 감성사전 대비 맞춤형 감성사전의 예측정확도가 통계적으로 유의한 수준의 성능 향상을 보였다. 데이터 영역의 특성에 맞는 맞춤형 사전 구축을 통한 감성분석의 예측의 성능 향상을 확인하였다.

윤환중(2015)은 감성 분석 알고리즘에 대한 연구가 부족하고 감성 사전의 구축이 미비하여 대부분 기계학습을 통한 감성 분석에 치중하고 있기 때문에 한국어 문법에 기초한 감성 분석 알고리즘의 연구는 미흡하고 부족한 실정을 고려하여 트위터 데이터의 활용과 한국어 감성 분석을 위한 알

고리즘 및 감성분석 시스템을 제안하였다. 극성 분석을 위하여 한국어 감성 사전의 구축이 요구되었기에, SentiWordNet이라는 영어 감성사전의 데이터를 추출하여 제시한 알고리즘에 맞게 재구성하고 경량화 한 뒤 한국어로 번역하여 특정 분야에 특화되지 않은 범용적인 한국어 감성 사전을 구축하였다.

성진, 원규식, 이수원(2015)은 Web Crawler를 통해 2014 브라질월드컵과 관련된 웹 뉴스 댓글을 수집하여 형태소분석기를 이용해 전처리를 실시하였으며, 한국어 감성사전을 기반으로 브라질 월드컵 관련 감성단어 및 신조어를 추가하여 스포츠 감성사전을 구축하여 브라질 월드컵의 주요 이슈를 탐지하고, 한국 축구대표팀 관련 인물에 대한 감성을 분석하였다.

2.2 감성분석을 위한 머신러닝 알고리즘 소개

감성분석을 위한 머신러닝 알고리즘은 서포트 벡터 머신(Support Vector Machine)과 나이브 베이즈(Naive Bayes)를 기반으로 연구되고 있다.

서포트 벡터 머신은 기계학습의 하나이다. Vladimir Naumovich Vapnik에 의해 1998년에 제안된 통계적 학습이론으로서 패턴 인식과 자료 분석을 위한 지도 학습 모델이며, 주로 분류와 회귀 분석을 위해 널리 사용한다.

서포트 벡터 머신의 기본 원리는 훈련 데이터들을 고차원의 특징공간으로 사상(Mapping)시킨 후에 두 분류 사이의 여백(Margin)을 최대화 시키는 결정함수(Hyperplane)를 찾는 것이다. 서포트 벡터 머신은 이론적 근거에 명확하여 두고 있어 결과 해석이 쉽고, 실제 응용함에 있어서 인공지능 경망 수준의 높은 성과를 내며, 적은 학습 자료만으로 신속하게 분별 학습을 수행하기 때문에 여러 분야에서 사용하고 있다. 그림 2는 서포트 벡터 머신의 기본 원리를 소개한다.

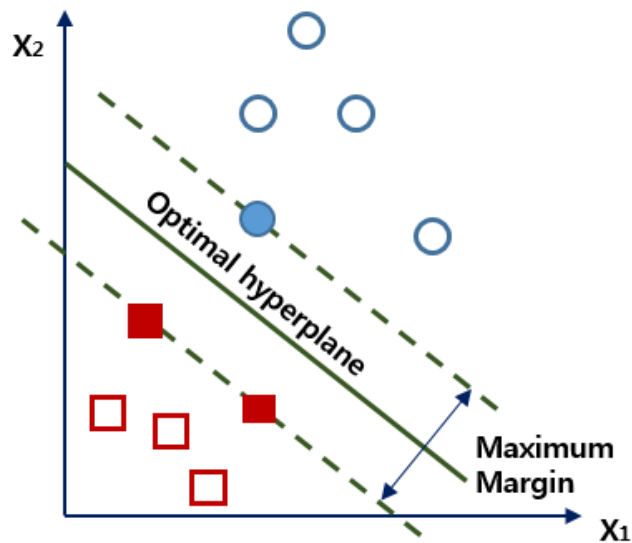


그림2. 서포트 벡터 머신

나이브 베이즈 분류기는 조건부 확률에 베이즈 정리(Bayes Theorem)를 적용하고 문서나 데이터를 구성하는 각각의 요소들이 등장할 확률에 대

한 독립성을 가정하여 입력 벡터를 분류하는 확률적 분류이다. 나이브 베이즈 알고리즘은 일반적으로 정확도가 높은 편이며, 알고리즘의 구조와 가정이 비교적 단순하지만 실제 상황에서 매우 잘 작동한다. 또한 학습데이터의 양이 많을 때 더욱 정확도가 높다. 1950대 이후 광범위하게 연구되고 있으며, 적절한 전처리를 거치면 서포트 벡터 머신과도 경쟁할 만큼 우수한 분류 성능을 보여준다.

2.3 머신러닝 알고리즘을 이용한 감성분석 사례

김동균, 허지용, 조지훈, 박수영, 김용혁(2011)은 사람의 다양한 감정을 분류하는 방법으로 기계학습을 사용하였다. 트위터 사용자의 감정을 나타내는 트위터 데이터를 구성하는 단어들을 사용하여 기계학습 후 트위터 데이터를 작성한 사용자의 감정을 분류하였다. 기계학습 중에서도 지도학습인 서포트 벡터 머신 알고리즘을 이용하여 트위터 데이터의 감정 유무를 분류하고, 감정이 있는 트위터 데이터의 경우 4가지의 감정으로 분류하였다. 그 결과 평균적으로 약 75%의 적중률로 실제 감정을 파악함에 있어 성공적인 결과를 확인하였다.

박종호(2015)는 2014년 서울시장 선거에서 수집된 약 17만 건의 트위터 데이터 중, 3121건의 샘플 데이터로 극성 예측 모델을 생성하고, 이 예측 모델을 통하여 새로운 트위터 데이터의 극성을 자동으로 예측하는 연구

를 진행하였다. 2014년 3월 9일부터 5월 18일까지의 서울시장 선거 트위터
터를 수집하여, 각 후보였던 “박원순”, “정몽준” 키워드로 데이터를 수
집하여 수집된 데이터에 대하여 기계학습의 일종인 서포트 벡터 머신을 통
한 극성예측모형을 도출하였다. 이 연구를 통하여, 선거데이터와 같은 극
성분석이 명확한 데이터는 서포트 벡터 머신과 같은 모형을 통하여, 자동
으로 새로운 트위터의 극성을 판단할 수 있는 것으로 분석하였다.

송은지(2015)는 온라인 웹사이트에서 수집한 데이터는 띄어쓰기와 철자
등에 오류가 많아 기존의 사용하던 형태소 분석기로는 정확한 분석을 할
수는 없는 문제점을 해결하기 위해서 초성과 중성, 어절 패턴 사전을 이용
해서 보정할 수 있는 감성분석 모듈을 제안하였다. 감성분석 모듈구성에
있어 학습에 의해 누적된 속성DB를 사용하여 긍정과 부정의 글들을 분류
하도록 하였고 이진 분류 시스템 중 가장 효율적인 서포트 벡터 머신 알고
리즘을 사용하였다.

제 3 장 연구 방법

3.1 데이터 수집 및 전처리 방법

본 연구에서는 20대 국회의원 선거의 트위터 데이터 감성분석을 위하여 트위터 API인 Twitter4J의 Java Wrapper를 사용하여 20대 국회의원 선거의 공식 선거운동 기간인 2016년 3월 31일부터 2016년 4월 12일까지 "총선", "국회의원", "선거", "413", "새누리당", "더불어민주당", "더민주", "새누리", "국민의당", "공천" 등 10개의 키워드로 총 1,048,574건의 트위터 데이터를 수집하였다. 그림 3은 트위터 API를 이용하여 수집한 데이터를 보여준다.

| | |
|--------|--|
| 621279 | 아닌건 아니다. 틀린건 틀린거다. 옳지 않은건 옳지 않은것이다. 나쁜놈은 나쁜놈이다. |
| 621280 | 진실과 정의사회, 경제 민주화, 양극화 해소,복지사회 #SNS소통(RT전달자) #인성교육이 절실히 요구되는 사회, 상식이 통하는 사람사는세: |
| 621281 | 두아이의 아빠/아이들이 더 나은 나라에서 살기를 바램/안철수/김성식/국민의당/정책네트워크내일 |
| 621282 | 정의가살아숨쉬는 |
| 621283 | 친일청산! 경제민주화! 평화통일! (광고계정.명연계정과 트윗없는 급조된 계정은 알바로 간주합니다.야당의 특정정치인 알바들도 팔로우 하 |
| 621284 | 나는 '나비, 사바나로 날다' '엄마야 누나야 강변살자'를 쓴 작가다. 사람들이 좋고 그 사람들 이야기는 더 좋다. 사랑스런 사람들아 모여라. i |
| 621285 | 진DJ진노진문~♡[사람이 먼저다] |
| 621286 | 우주의 끝은? / RT 사절입니다./인용RT사절입니다./깨시,수줍사절/담대한 변화가 시작됩니다./국민의당 |
| 621287 | 선거때마다 국민의당 아니면 정의당 후보에게만 희생 강요 하는지 모르겠다. 그리고 더민주 후보는 한번도 희생한 적없다. 다른 당 후보에? |
| 621288 | seogyo-dong, mapping planner |
| 621289 | 총선 유권자로서도 올이다π#마포율713901676856344576 |
| 621290 | 머리카락 엉덩이까지 기를 수 있는 꿈이이루어진, 실크로드를 가로질러 서유령을 가고 싶은 사람 |
| 621291 | 이나라에 정의가 살아숨쉬는 그날까지~언론이 언론다워야 정의가 바로선다. |
| 621292 | 노무현의 정의를 사랑한다.그의 죽음까지도 사랑한다용기있고 결단할 수 있는 사람만이 죽음도 겨안을 수 있기에...이름만 들어도 늘 가슴이 |
| 621293 | 어른의 순수는지혜에서 나온다. |
| 621294 | 비어지는 공간을 보고 슬퍼하지마라 채워지는 가슴을 기뻐하라. 정의없는 힘을 배격한다! 자신을 비난하지 마라 당신의 잘못이 아니다 / 노 |
| 621295 | 새누리 "150석+α" 더민주 "130석" 국민의당 "20석" 다음 뉴스 https://t.co/P9cUJ6uWF7 아주 소설을 쓰는구나.. 연합아 똥이나 퍼먹어라!!7 |
| 621296 | 문재인 대표님이 영입한 인재들이 험지에서 뛰고 있습니다. 적극 도와서 총선 승리합시다 |
| 621297 | 지못미는 한 번으로 족하다! 그래서 문대표를 지켜야 한다! 친노는 자랑해야할 별명이다! 그래서문빠여야 한다! |
| 621298 | 샐러리맨의 고단한 하루 그래도내일의 희망을 갖고 오늘도 출근한다.. |
| 621299 | 수제쿠키 앤 머핀굽는 여자/아이들을 사랑하는 아동요리강사/쿠키를 통하여 따뜻한 사람나누기~♡♡♡ |
| 621300 | 죽기 전에 꼭 해야할 10가지, 표상하고 개입하기 |
| 621301 | 시급 올리고 월급 올리고 연봉 올리려면 무조건 문재인대통령 만들 수 있는 총선 결과 만들어야 가능해요 과반 달성 못하면 개불도 결국 안 |
| 621302 | 상식... |
| 621303 | 신월동학박이/달바보/국민TV조합원/더불어민주당 당원/팩트TV후원(약소)/대머리/마음글에 제가 좋아하는 노래 되게많아요/거북이 두 마리: |
| 621304 | 진보하라 역사를 바로 알라 |
| 621305 | 대한민국 4월 총선 부정선거 - 개표 조작 감시해야 합니다. |

그림 3. 트위터 API를 이용한 데이터 수집

수집된 총 1,048,574건의 트위터 데이터를 한나눔 형태소 분석기(Java)를 활용하여 관사, 전치사, 조사, 접속사와 같은 불용어를 제거하고 트위터 데이터의 원문을 각각 형태소 단위로 분리하였다. 이를 인해 형태소 단위로 분리된 하나 하나의 단어를 독립변수로 설정하였다. 그림 4는 트위터 데이터를 형태소 단위로 분리한 결과를 보여준다.

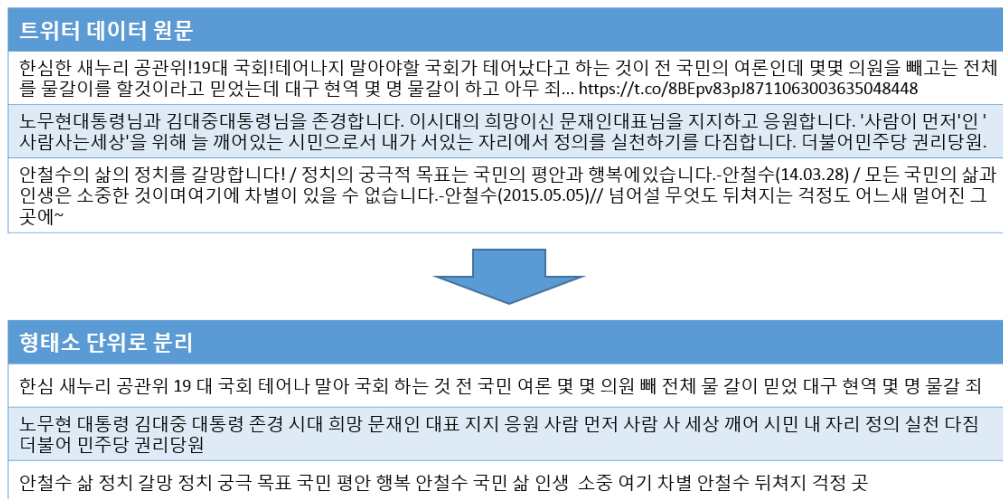


그림 4. 트위터 데이터 형태소 분리

형태소 단위로 분리된 트위터 데이터의 개별의견에 극성변수를 추가하는 극성 태깅을 실시 하였다. 극성 변수는 이진형 변수로 구성되며, 해당 의견이 대상에 긍정적인 경우에는 1, 부정적인 경우에는 0의 값을 할당하여 각각의 트위터 데이터 마다 극성을 표시하였다.

본 연구에서는 수집된 20대 국회의원 선거 관련 트위터 데이터에서 “새

누리당”, “더불어민주당”, “국민의당”에 관하여 트위터 데이터의 극성을 태
 기하였다. 각 정당 별로 독립적으로 수행하였으며, 트위터의 의견이 해당
 정당에 긍정적인 영향을 미칠 것으로 예상되는 경우에는 1, 부정적인 영향
 을 미칠 것으로 예상되는 경우에는 0으로 총 5,559개의 트위터에 대하여
 극성을 표시하였다. 그 결과 새누리당의 경우 총 1,126개, 더불어민주당의
 경우 총 4,056개, 국민의당의 경우에는 747개의 트위터 데이터에 대해서
 각각 극성이 부여되었다. 그림 5는 트위터 데이터의 극성 태기한 결과를
 보여준다.

| 형태소 단위로 분리된 트위터 데이터 | 새누리당 | 더불어민주당 | 국민의당 |
|---|------|--------|------|
| 공천학살 파동 여유 박근혜 새누리당 더민주 국민의당 야권 연대 공식 거 부 수도권 새누리당 압승 지상파 종편 편파방송 60대 새누리 지지층 압도 투표율 | 1 | 0 | 0 |
| 15년 역사 후퇴 반기문 반대 독재자 거수기 성누리당 반대 명분 탈당 철수 씨 반대 호남 민심 팔아 박지원 주승룡 박주선 박준영 안철수 전 정배 김한길 문병호 김영환 조경태 반대 여론 왜곡 앞장 종편 거기 출 연 정치 평론가 들 반대 THE 민주 집결 | 0 | 1 | 0 |
| 안철수 지지자 국민의당 지지자 여러분 께 또 간곡 호소 합니다 지금 총 선 기간 엄중 시간 새누리 친박 공천 독재 더민당 친문 김종인 독재 공천 김종인 셀프 비례 2번 공천 1번 논문 표절 국민의당 우주 기운 당 내 상호 총질 중단 단결 | 0 | 0 | 1 |

그림 5. 트위터 데이터의 극성 태기

극성 태기된 트위터 데이터에 대하여 각각의 단어를 하나의 독립변수로,
 극성변수인 “새누리당”, “더불어민주당”, “국민의당”을 종속변수로 하는 데
 이터 행렬을 구성하였다. 각각의 트위터에 포함되는 단어는 1, 포함되지
 않는 단어는 0을 부여하였으며, 각 정당 별로 극성변수를 변경해가며 총
 세 종류의 데이터 행렬을 생성하였다. 그림 6은 트위터 데이터의 행렬화

결과를 보여준다.

| Tweet No. | 독립변수 | | | | | 종속변수 | |
|-----------|------|--------|------|-----|-----|------|-----------|
| | 새누리당 | 더불어민주당 | 국민의당 | 개헌선 | ... | 박근혜 | 극성변수_새누리당 |
| 1 | 0 | 1 | 0 | 0 | ... | 0 | 0 |
| 2 | 0 | 0 | 0 | 1 | ... | 0 | 0 |
| 3 | 1 | 0 | 0 | 1 | ... | 0 | 0 |
| 4 | 1 | 0 | 0 | 0 | ... | 1 | 1 |
| 5 | 0 | 0 | 0 | 1 | ... | 0 | 0 |
| 6 | 0 | 0 | 1 | 0 | ... | 0 | 0 |
| 7 | 0 | 1 | 0 | 0 | ... | 1 | 1 |
| 8 | 1 | 0 | 0 | 0 | ... | 0 | 1 |

그림 6. 트위터 데이터의 행렬화

3.2 Lasso 및 Ridge 선형회귀 알고리즘

Lasso 회귀는 Tibshirani(1996)에 의해 제안된 방법론으로 회귀 계수의 축소를 통하여 예측 정확도를 높이고, 이와 함께 영향력이 적은 회귀계수 값을 쉽게 0으로 수렴하게 만드는 변수선택의 기능을 가지고 있어 해석력을 높여준다. Lasso 회귀는 Ridge 회귀의 예측 정확도와 변수선택의 해석력을 모두 가진 분석 방법론으로 널리 알려져 있다. Lasso 회귀의 추정량은 다음의 식(1)과 같이 구할 수 있다.

$$\beta^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (1)$$

Lasso 회귀에서는 회귀계수의 절대값의 합을 페널티(Penalty)로 대입한다. 한편 λ 는 tuning parameter가 크면 클수록 보다 많은 회귀계수를 0으로 수렴한다. Lasso 회귀는 λ 제약 조건을 사용하여 영향력이 없는 변수의 회귀계수를 0으로 만들어 예측에 필요한 중요한 변수만을 선택해 차원을 축소하고, 변수 선택이 가능해 예측모형의 해석력을 높여주며, 회귀계수의 절대값에 페널티를 부여하기 때문에 적은 수의 변수만을 선택하여 해석을 편리하게 한다는 장점을 가진다.

Ridge 회귀는 Hoerl와 Kennard(1971)가 제안한 방법론으로 회귀계수의 크기에 페널티를 부여함으로써 회귀계수를 축소하는 방법이다.

Ridge 회귀에서는 회귀계수의 절대값의 합을 페널티로 대입한다. 한편 λ 는 tuning parameter가 크면 클수록 보다 많은 회귀계수를 0으로 수렴한다. Ridge 회귀의 추정량은 다음의 식(2)와 같이 구할 수 있다.

$$\beta^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (2)$$

식(2)의 $\lambda \geq 0$ 은 회귀계수 크기의 축소 양을 결정하는데 λ 의 값이 커질수록 축소의 양이 많아지며 회귀계수가 0에 가까워진다.

Ridge 회귀분석은 λ 제약 조건을 이용하여 회귀계수의 크기를 축소함으로써 다중공선성의 문제를 해결하고, 예측정확도를 높인 방법이다. Ridge

회귀는 조율모수에 의해 최소제곱 추정치보다 회귀계수 크기가 축소되어 편의는 있지만 분산을 줄이기 때문에 예측오차를 줄일 수 있다는 장점이 있다.

Lasso 및 Ridge 회귀 분석의 차이점은 각각 다른 변수 감소 방법을 사용한다는 것이다. Lasso의 경우 중요한 변수는 유지되지만 나머지는 제외된다. Ridge 회귀에서는 중요하지 않은 변수의 계수를 0에 가까운 값으로 줄일 수는 있지만 해당 변수를 0으로 만들면 불완전하게 제거 할 수 없다.

3.2 검증방법

본 연구에서는 개별 트위터의 극성을 분류하기 위해 Lasso 및 Ridge 회귀 모형을 이용한 감성분석을 수행하였다. 극성변수에 큰 영향을 미치는 독립변수를 추출함과 동시에 주요 독립변수가 극성변수에 미치는 극성을 평가하고 학습된 Lasso 및 Ridge 회귀 모형을 이용해 새로운 트위터 데이터의 극성을 예측하였다.

주요 독립변수인 트위터 내의 단어의 추출 및 극성 분석을 위한 Lasso 및 Ridge 회귀 모형에 대해 식 (3)과 (4)의 목적 함수를 만족하는 회귀 계수를 각각 추정한다. 여기에서 β_j 는 회귀 계수, λ 은 회귀계수 β_j 의 크기를 결정하는 파라미터, p 는 독립 변수의 수를 나타낸다.

$$\text{Minimize } l(\beta_0, \beta_1, \dots, \beta_p) + \lambda \sum_{j=1}^p |\beta_j| \quad (3)$$

$$\text{Minimize } l(\beta_0, \beta_1, \dots, \beta_p) + \lambda \sum_{j=1}^p \beta_j^2 \quad (4)$$

이진 값 응답의 경우 손실 함수는 식 (5)와 (6) (Meier, Van De Geer, & Bühlmann, 2008)과 같이 정의된다. 여기서 x_{ij} 는 독립 변수이고 i 번째 데이터의 j 번째 독립변수를 의미한다. y_i 는 종속 변수이고 i 번째 데이터의 극성변수를 나타내며 n 은 데이터 양을 나타낸다.

$$l(\beta_0, \beta_1, \dots, \beta_p) = - \sum_{i=1}^n y_i \times \log p(x_i) + (1 - y_i) \times \log(1 - p(x_i)) \quad (5)$$

$$p(x_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \quad (6)$$

식 (5)에서 $p(x_i)$ 는 데이터가 긍정일 확률을 의미. $y_i = 1$ 일 때 즉, 트위터 데이터가 긍정이면, 식 (5)는 $\log p(x_i)$ 로 감소되고 $p(x_i) = 1$ 에서 최소화된다. $y_i = 0$ 일 때 즉, 트위터 데이터가 부정인 경우, 식 (5)는 $\log(1 - p(x_i))$ 가 되며, $p(x_i) = 0$ 에서 최소화 된다.

따라서 Lasso와 Ridge 회귀 모델은 긍정 또는 부정의 트위터 데이터가 정확하게 분류되어 손실 함수를 최소화하도록 훈련된다. Lasso와 Ridge

회귀 모형 학습에서 사용되지 않는 새로운 트위터 데이터의 경우, 학습된 모델은 태그가 붙은 트위터 데이터를 입력으로 받아들이고 종속 변수의 출력 값을 사용하여 트위터 데이터의 극성을 감지한다.

아래 그림 7과 같이 트위터 데이터의 극성분석을 위한 Lasso 및 Ridge 회귀 모형 수행코드를 작성하였다.

```
label<-tw$새누리당)
label2<-label[!is.na(label)]
sa_data<-sa_data[!is.na(label),]

sanuri_label<-label2[!label2==0] ## 중립 라벨 제외
sanuri_label<-ifelse(sanuri_label==1,1,0) ## 긍정 =1, 부정 =0 으로 코딩
table(sanuri_label)
sanuri_data<-sa_data[!label2==0,]
sam<-sample(1:length(sanuri_label),length(sanuri_label)*0.7)
fit2=glmnet(as.matrix(sanuri_data[sam,]),sanuri_label[sam],family="binomial",alpha=1) ### Lasso 모델링

dat<-coef(fit2,s=0.001) #
str(dat)
posi_selec_var<-matrix(dat)[ matrix(dat) >0] ## 양의 계수 저장
posi_naems<-rownames(dat)[ matrix(dat) >0]
names(posi_selec_var)<-posi_naems
a<-posi_selec_var[order(-posi_selec_var)] ## 계수가 높은 순으로 정렬
a<-data.frame(a)
write.csv(a,"positive_lasso_새누리당.csv") ## 계수 저장

posi_selec_var[order(posi_selec_var)][1:100]

nego_selec_var<-matrix(dat)[ matrix(dat) <0] ## 음의 계수 저장
nego_naems<-rownames(dat)[ matrix(dat) <0]
names(nego_selec_var)<-nego_naems
nego_selec_var[order(nego_selec_var)][1:100] ## 계수가 낮은 순으로 정렬
b<-nego_selec_var[order(nego_selec_var)]
b<-data.frame(b)
write.csv(b,"negative_lasso_새누리당.csv") ## 계수 저장

pre<-predict(fit2,newx=as.matrix(sanuri_data[-sam,]),s=seq(0.0001,0.001,0.0001)) # make predictions
pred<-ifelse(pre[,1] > 0,1,0)
sum(pred == sanuri_label[-sam])/length(pred)
test_y<-sanuri_label[-sam]
sum(test_y == 1 & pred == 1)/sum(test_y == 1)
sum(test_y == 0 & pred == 0)/sum(test_y == 0)
```

그림 7. Lasso 및 Ridge 회귀 수행코드

제 4 장 실험 결과 및 분석

본 장에서는 트위터 데이터의 감성분석을 위하여 Lasso 및 Ridge 회귀 분석 알고리즘을 통해 각 정당에 대한 트위터 데이터의 감성분석 결과 및 예측 정확도 그리고 도출된 결과에 대한 분석을 수행한다.

4.1 20대 국회의원 선거 트위터 데이터

20대 국회의원 선거 정당 지지율 예측을 위한 트위터 감성 분석을 위하여 공식선거운동 기간인 2016년 3월 31일부터 2016년 4월 12일까지 선거 관련 수집된 데이터 1,048,574건 중 5,559건의 데이터를 샘플링하여 데이터 전처리 및 극성태깅 한 결과 트위터의 극성 분포는 전반적으로 각 정당에 대한 부정적인 트위터 데이터가 긍정적인 트위터 데이터보다 많은 것을 보여준다.

새누리당의 경우 전체 1,126건의 트위터 데이터 중 긍정적인 영향을 미치는 트위터 데이터는 18.3%, 부정적인 영향을 미치는 트위터 데이터는 81.7%이며, 더불어민주당의 경우 전체 4,061건의 트위터 데이터 중 긍정적인 영향을 미치는 트위터 데이터는 20.9%, 부정적인 영향을 미치는 트위터 데이터는 79.1%로 나타났다. 또한, 국민의당의 경우 전체 747건 중 긍정적인 영향을 미치는 트위터 데이터는 46.1%, 부정적인 영향을 미치는

주는 트위터 데이터는 53.9%로 나타났다. 이러한 트위터 데이터의 분포를 분석한 결과 20대 국회의원 선거 관련 트위터 데이터는 각 정당에 긍정적인 영향보다는 부정적인 영향을 미치는 것으로 나타났다. 그림 8은 각 정당별 트위터 데이터의 긍정과 부정의 분포를 보여준다.

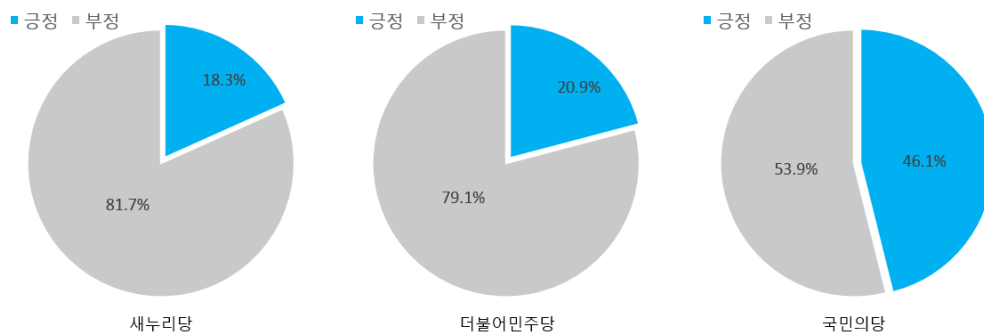


그림 8. 각 정당별 트위터 데이터 분포

4.2 트위터 감성분석 결과

20대 국회의원 선거 정당 지지율 예측을 위한 트위터 감성 분석을 위하여 2016년 3월 31일부터 2016년 4월 12일까지 선거 관련 수집된 트위터 데이터 1,048,574건 중 5,559건을 샘플링하여 데이터 전처리 및 극성 태깅, Lasso & Ridge 회귀 모형 도출을 통한 감성분석 결과이다.

새누리당의 경우 총 1,126개, 더불어민주당의 경우 총 4,061개, 국민의당의 경우에는 747개의 트위터 데이터에 대해서 각각 극성이 부여되었으

며, 이를 통해 생성된 모형과 예측모형의 극성분석 정확도를 검증하였다.

4.2.1 새누리당 결과 및 해석

다음은 20대 국회의원 선거 관련 트위터 데이터 중 극성 태깅된 전체 5,559개의 중 새누리당에 대한 1,126건의 트위터 데이터에 대한 Lasso 및 Ridge 회귀 모형 도출을 통한 감성분석 결과와 정확도를 나타낸 결과이다.

아래의 표는 새누리당에 대한 트위터 데이터 1,126건의 감성 분석 정확도의 결과이다.

표 1. 새누리당 트위터 감성분석 정확도 결과

| | 전체 정분류율 | 긍정 트윗 정분류율 | 부정 트윗 정분류율 | 극성표시건수 (긍정/부정) |
|-------|------------|---------------|---------------|-------------------|
| Lasso | 0.98 | 1 | 0.97 | 207/919 |
| Ridge | 0.908 | 0.52 | 1 | 207/919 |

(lamda 0.001)

Lasso 회귀 모형의 경우 전체 정분류율 0.98, 긍정 트윗정 분류율 1, 부정 트윗 정분류율 0.97이며, Ridge 회귀 모형의 경우 전체 정분류율 0.908, 긍정 트윗 정분류율 0.52, 부정 트윗 정분류율 1로써 Lasso 회귀 모형이 Ridge 회귀 모형보다 트위터 감성분석의 정확도가 높다는 것을 나

타낸다. 결과적으로 Lasso 회귀 모형이 Ridge 회귀 모형보다 트위터 극성 분석에서 더 나은 성능을 보여주며, 여론의 극성을 더 잘 반영한다.

새누리당의 부정 트윗 수가 긍정 트윗 수보다 월등히 많은 것은 당시 민심이 새누리당에 부정적이었다는 여론을 반영한 것으로 판단된다. Ridge 회귀 모형에서 부정 트윗 분류율은 높고 긍정 트윗 분류율이 낮은 것은 긍정 트윗 수가 상당히 적어 학습데이터로써 긍정 데이터 수가 상대적으로 적어 학습을 통해서 긍정은 무시하고 부정의 분류율을 높이려고 하는 Lasso 및 Ridge 모형의 클래스 불균형 문제(Class Imbalance Problem)로 인하여 나타난 결과이다.

아래의 표는 새누리당에 대한 Lasso 회귀 모델링 결과 긍정적인 영향을 미친 주요 단어를 정리한 표이다.

표 2. 새누리당 Lasso 결과 긍정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|------|----------|------|----------|---------|----------|
| 역적들 | 11.09673 | 다짐 | 9.981597 | 비상대책위원회 | 9.634942 |
| 야권연대 | 8.472913 | 김용민 | 8.277692 | 강동원 | 8.275951 |
| 민주적 | 7.92351 | 비전 | 7.827175 | 영구집권 | 7.768282 |
| 김용민 | 8.277692 | 강동원 | 8.275951 | 이혜훈이 | 7.475086 |
| 이한성 | 7.358739 | 대통령님 | 6.881949 | 서구 | 6.711004 |
| 기호1번 | 6.37658 | 개헌선 | 6.325329 | 전현희 | 5.955374 |
| 구로 | 5.937658 | 추대 | 5.772347 | 전하 | 5.684004 |
| 김종인표 | 5.628305 | 천노 | 5.02147 | 강석훈 | 4.996708 |

| | | | | | |
|-----|----------|----|----------|-------|----------|
| 투표 | 3.394661 | 서초 | 5.772347 | 민주당 | 3.240243 |
| 윤상현 | 3.189823 | 우과 | 2.932034 | 박대통령을 | 2.764312 |

새누리당에 긍정적인 영향을 미친 주요 단어는 “야권연대”, “김용민”, “영구집권”, “대통령님”, “서구”, “친노” 등이 볼 수 있으며 당시의 일반 민심이 새누리당에 우호적이지 못했기에 전통적인 새누리당 지지층이 결집한 것으로 보인다.

아래의 표는 새누리당에 대한 Lasso 회귀 모델링 결과 부정적인 영향을 미친 주요 단어를 정리한 표이다.

표 3. 새누리당 Lasso 결과 부정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|-------|----------|------|----------|--------|----------|
| 김어준 | -7.11503 | 경선탈락 | -6.31543 | 비례1번으로 | -4.09367 |
| 노무현정부 | -4.00251 | 세월호 | -3.662 | 노망난 | -3.62573 |
| 대리인 | -3.27883 | 박주민 | -2.92423 | 청와대 | -2.69336 |
| 민집모 | -2.59005 | 인재영입 | -2.521 | 5년 | -2.37307 |
| 이한구 | -1.79404 | 선대위 | -1.78951 | 성완중 | -1.66946 |
| 김종인 | -1.51276 | 원내대표 | -1.50021 | 집권 | -1.4924 |
| 국정원 | -1.44266 | 막장 | -1.41062 | 진박 | -1.40919 |
| FTA | -1.39912 | 표창원 | -1.38497 | 대권 | -1.36792 |
| 이상돈 | -1.26317 | 나경원 | -1.20095 | 이명박구속 | -1.18676 |
| 박근혜 | -1.15884 | 심판 | -1.15704 | 단죄 | -1.12152 |

새누리당에 부정적인 영향을 미친 주요 단어를 보면 “노무현정부”, “세월호”, “성완중”, “김종인”, “국정원”, “진박”, “FTA” 등의 단어가 보인다. 세월호 사건이 20대 국회의원 선거에도 새누리당에 부정적인 영향을 준 것으로 보이며, 당시 공천과정에서 문제가 되었던 “진박”은 새누리당에 부정적인 영향을 준 것으로 보인다.

아래의 표는 새누리당에 대한 Ridge 회귀 모델링의 결과 긍정적인 영향을 미친 주요 단어를 정리한 표이다.

표 4. 새누리당 Ridge 결과 긍정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|------|----------|-------|----------|------|----------|
| 개헌선 | 0.733294 | 남구 | 0.71998 | 추대 | 0.705077 |
| 최명길 | 0.655788 | 살길 | 0.645978 | 전권 | 0.634021 |
| 구도 | 0.632107 | 트위터 | 0.610489 | 역적들 | 0.561641 |
| 전희경 | 0.558944 | 계열 | 0.555805 | 180석 | 0.553044 |
| 강석훈 | 0.543965 | 측근 | 0.542285 | 행보 | 0.542284 |
| 종편에서 | 0.541134 | 음모 | 0.539912 | 철퇴 | 0.53991 |
| 최고위 | 0.513531 | 1위 | 0.491844 | 200석 | 0.462702 |
| 역선택 | 0.457499 | 현상황 | 0.457495 | 숫자 | 0.455959 |
| 중랑갑 | 0.454824 | 애국 | 0.454051 | 대통령님 | 0.450886 |
| 이혜훈이 | 0.441873 | 노무현재단 | 0.438039 | 민변 | 0.43802 |

아래의 표는 새누리당에 대한 Ridge 회귀 모델링의 결과 부정적인 영향을 미친 주요 단어를 정리한 표이다.

표 5. 새누리당 Ridge 결과 부정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|---------|----------|------|----------|-------|----------|
| 협박 | -0.41292 | 약속 | -0.27506 | 김헌태 | -0.19201 |
| 2016 | -0.19164 | 교감 | -0.19102 | 김종인 | -0.19057 |
| 당대표 | -0.19057 | 청년들 | -0.19032 | 공천권 | -0.18685 |
| 세월호망치부인 | -0.18618 | 유승민계 | -0.18478 | 조웅천 | -0.18222 |
| 유세 | -0.18111 | 한계 | -0.17998 | 강동원 | -0.17979 |
| 아수라장 | -0.17963 | 과당 | -0.17660 | 특정인 | -0.17660 |
| 서청원 | -0.17472 | 알바들 | -0.17315 | 선언 | -0.17243 |
| 장담 | -0.17201 | 철수당 | -0.17141 | 내치 | -0.17082 |
| 2중대가 | -0.16974 | 김용민 | -0.16902 | 김어준 | -0.16902 |
| 200명 | -0.16797 | 단죄 | -0.16601 | 민족반역자 | -0.16561 |

4.2.2 더불어민주당 결과 및 해석

다음은 20대 국회의원 선거 관련 트위터 데이터 중 극성 태깅된 전체 5,559개의 중 더불어민주당에 대한 4,601건의 트위터 데이터에 대한 대한 Lasso 및 Ridge 회귀 모형 도출을 통한 감성분석 결과와 정확도를 나

타낸 결과이다

아래의 표는 더불어민주당에 대한 트위터 데이터 4,911건의 감성 분석 정확도의 결과이다.

표 6. 더불어민주당 트위터 감성분석 정확도 결과

| | 전체 정분류율 | 긍정 트윗 정분류율 | 부정 트윗 정분류율 | 극성표시건수 (긍정/부정) |
|-------|------------|---------------|---------------|-------------------|
| Lasso | 0.98 | 1 | 0.98 | 850/3211 |
| Ridge | 0.83 | 0.21 | 1 | 850/3211 |

(lamda 0.001)

Lasso 회귀 모형의 경우 전체 정분류율 0.98, 긍정 트윗정 분류율 1, 부정 트윗 정분류율 0.98이며, Ridge 회귀 모형의 경우 전체 정분류율 0.83, 긍정 트윗 정분류율 0.21, 부정 트윗 정분류율 1로써 Lasso 회귀 모형이 Ridge 회귀 모형보다 트위터 감성분석의 정확도가 높다는 것을 나타낸다. 결과적으로 Lasso 회귀 모형이 Ridge 회귀 모형보다 트위터 극성 분석에서 더 나은 성능을 보여주며, 여론의 극성을 더 잘 반영한다.

더불어민주당의 부정 트윗 수가 긍정 트윗 수보다 월등히 많은 것은 당시 민심이 더불어민주당에 부정적이었다는 여론을 반영한 것으로 판단된다. Ridge 회귀 모형에서 부정 트윗 분류율은 높고 긍정 트윗 분류율이 낮은 것은 긍정 트윗 수가 상당히 적어 학습데이터로써 긍정 데이터 수가 상대

적으로 적어 학습을 통해서 긍정은 무시하고 부정의 분류율을 높이려고 하는 Lasso 및 Ridge 모형의 클래스 불균형 문제로 인하여 나타난 결과이다.

아래의 표는 더불어민주당에 대한 Lasso 회귀 모델링 결과 긍정적인 영향을 미친 주요 단어를 정리한 표이다.

표 7. 더불어민주당 Lasso 결과 긍정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|------|----------|------|----------|-----------------|----------|
| 해산 | 12.28003 | 밀실공천 | 10.62646 | 천년만년 | 9.885214 |
| 검찰 | 9.179856 | 개혁적 | 8.892712 | 색누리 | 8.760529 |
| 심판하자 | 8.641303 | 용기 | 8.454734 | 도종환 | 7.493325 |
| 공천권 | 7.421495 | 집결 | 7.419973 | 분탕질 | 7.097411 |
| 진영의원 | 6.385993 | 후폭풍 | 6.323978 | 집권당 | 5.551765 |
| 견제 | 5.496658 | 정치보복 | 5.473092 | 더불어민주당 권리 당원 | 5.223524 |
| 은평갑 | 5.174573 | 이삭줍기 | 5.085229 | 150석 | 5.006941 |
| 조웅천 | 4.986889 | 노회찬 | 4.816273 | 보수층 | 4.703535 |
| 정치꾼 | 4.656715 | 친노무현 | 4.600887 | 김용환 | 4.577005 |
| 60대 | 4.393906 | 탈당파 | 4.357004 | 손혜원님 | 4.05455 |

더불어민주당에 긍정적인 영향을 준 주요 단어를 보면 “개혁적”, “색누리”, “심판하자”, “집결”, “집권당”, “분탕질” 등의 단어가 보인다. 당시 집권 여당에 대한 심판과 개혁을 원하는 민심이 반영된 것으로 보인다.

아래의 표는 더불어민주당에 대한 Lasso 회귀 모델링 결과 부정적인 영향을 미친 주요 단어를 정리한 표이다.

표 8. 더불어민주당 Lasso 결과 부정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|-------|----------|------|----------|-----|----------|
| 접근금지 | -6.79952 | 광주시민 | -5.8332 | 서초 | -5.56877 |
| 바편애 | -5.38932 | 김무성계 | -4.98719 | 지도부 | -4.34521 |
| 민중연합당 | -3.77862 | 강동원 | -3.7397 | 침몰 | -3.52503 |
| 양비론 | -3.45963 | 퇴출 | -3.35982 | 여자 | -3.19606 |
| 부산 | -3.14268 | 비리 | -3.12051 | 여당 | -3.06945 |
| 유승민의원 | -2.97589 | 아사리판 | -2.94595 | 투표율 | -2.83911 |
| 김종인씨 | -2.73768 | 멋대로 | -2.72368 | 내홍 | -2.65775 |
| 국민의당 | -2.56064 | 밥그릇 | -2.53797 | 흡수저 | -2.47323 |
| 일베충 | -2.27709 | 목포 | -2.26614 | 개누리 | -2.23065 |
| 제자논문 | -2.22297 | 보수 | -2.14096 | 민생 | -2.11486 |

더불어민주당에 부정적인 영향을 미친 주요 단어를 보면 “광주시민”, “바편애”, “지도부”, “퇴출”, “김종인씨” 등의 단어가 보인다. 당시 광주지역에서 더불어민주당의 지지율이 좋지 않았으며, 지도부의 공천문제와 셀프공천이 부정적인 영향을 미친 것으로 보인다.

아래의 표는 더불어민주당에 대한 Ridge 회귀 모델링 결과 긍정적인 영향을 미친 주요 단어를 정리한 표이다.

표 9. 더불어민주당 Ridge 결과 긍정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|------|----------|----------------|----------|--------|----------|
| 눈높이 | 0.670461 | 용기 | 0.669807 | 자기당 | 0.658942 |
| 여왕 | 0.639741 | 정치보복 | 0.639641 | 친노패권주의 | 0.62149 |
| 경상 | 0.612143 | 남양주갑 | 0.607179 | 야당다운 | 0.598354 |
| 지도자 | 0.588209 | 진영의원 | 0.581184 | 망상 | 0.554199 |
| 상향식 | 0.550166 | 김재원 | 0.550153 | 도종환 | 0.549101 |
| 견제 | 0.545883 | 은평갑 | 0.537213 | 초선 | 0.535228 |
| 불복 | 0.531776 | 김제동 | 0.522311 | 노원병 | 0.514535 |
| 필승 | 0.497653 | 사람사는세상 | 0.487999 | 김경수 | 0.483836 |
| 집결 | 0.478280 | 더불어민주당 권리당원 | 0.478093 | 정봉주 | 0.473825 |
| 손혜원님 | 0.469141 | 패권주의 | 0.468700 | 현역의원 | 0.457535 |

아래의 표는 더불어민주당에 대한 Ridge 회귀 모델링 결과 부정적인 영향을 미친 주요 단어를 정리한 표이다.

표 10. 더불어민주당 Ridge 결과 부정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|-----|----------|------|----------|------|----------|
| 기권 | -0.23991 | 통합 | -0.20814 | 박주민 | -0.20049 |
| 갑질 | -0.19988 | 논문표절 | -0.18812 | 현역의원 | -0.18684 |
| 노망난 | -0.18312 | 2중대 | -0.17999 | 단죄 | -0.17845 |
| 퇴진 | -0.17467 | 빨갱이 | -0.17259 | 선대위 | -0.17106 |
| 대망론 | -0.17101 | 착각 | -0.17100 | 성소수 | -0.17046 |

| | | | | | |
|-------|----------|------|----------|-------|----------|
| 때려잡자 | -0.17046 | 조중동 | -0.17046 | 김종인에게 | -0.16927 |
| 폭망 | -0.16392 | 이구박퇴 | -0.16099 | 조웅천 | -0.15963 |
| 반발 | -0.15937 | 선거법 | -0.15826 | 박근혜퇴진 | -0.15804 |
| 경제민주화 | -0.15802 | 부정선거 | -0.15674 | 청년들 | -0.15574 |
| 김한길 | -0.15456 | 4대강 | -0.15252 | 탈당파 | -0.15085 |

4.2.3 국민의당 결과 및 해석

다음은 20대 국회의원 선거 관련 트위터 데이터 중 극성 태깅된 전체 5,559개의 중 국민의당에 대한 747건의 트위터 데이터에 대한 Lasso 및 Ridge 회귀 모형 도출을 통한 감성분석 결과와 정확도를 나타낸 결과이다.

아래의 표는 국민의당에 대한 트위터 데이터 4,911건의 감성 분석 정확도의 결과이다.

표 11. 국민의당 트위터 감성분석 정확도 결과

| | 전체 정분류율 | 긍정 트윗 정분류율 | 부정 트윗 정분류율 | 극성표시건수 (긍정/부정) |
|-------|------------|---------------|---------------|-------------------|
| Lasso | 0.99 | 1 | 0.98 | 345/402 |
| Ridge | 0.98 | 0.95 | 0.99 | 345/402 |

(lamda 0.001)

Lasso 회귀 모형의 경우 전체 정분류율 0.99, 긍정 트윗 정분류율 1, 부정 트윗 정분류율 0.98이며, Ridge 회귀 모형의 경우 전체 정분류율 0.98, 긍정 트윗 정분류율 0.95, 부정 트윗 정분류율 0.99로써 Lasso 회귀 모형이 Ridge 회귀 모형보다 트위터 감성분석의 정확도가 높다는 것을 나타낸다. 결과적으로 Lasso 회귀 모형이 Ridge 회귀 모형보다 트위터 극성 분석에서 더 나은 성능을 보여주며, 여론의 극성을 더 잘 반영한다.

국민의당의 경우 긍정 트윗 수와 부정 트윗 수가 비슷한 것은 당시 민심이 새로운 제 3정당에 대한 기대감으로 전반적인 여론이 우호적이었던 것으로 판단된다.

아래의 표는 국민의당에 대한 Lasso 회귀 모델링 결과 긍정적인 영향을 미친 주요 단어를 정리한 표이다.

표 12. 국민의당 Lasso 결과 긍정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|-------|----------|--------|----------|-------|----------|
| 비례명단 | 9.713208 | 서울시장 | 8.699945 | 공천학살 | 8.301786 |
| 국보위 | 6.729738 | 논문표절 | 6.678533 | 독선적 | 6.363161 |
| 국회의원 | 6.14132 | 숙의배심원단 | 5.229737 | 좌파 | 4.460918 |
| 노무현정신 | 4.452581 | 진영 | 4.164206 | 비대위 | 3.782702 |
| 당대표 | 3.599391 | 대선후보 | 3.462139 | 안철수님 | 3.388897 |
| 희망 | 3.274277 | 이명박정부 | 3.219111 | 김종인비례 | 3.164655 |
| 노원 | 3.101194 | 정권교체 | 3.021222 | 더불어 | 2.97585 |
| 지도부 | 2.905671 | 단일화 | 2.576054 | 공동대표 | 2.185599 |

| | | | | | |
|-----|----------|-------|----------|------|----------|
| 최측근 | 1.987788 | 당선권 | 1.897176 | 친박 | 1.849175 |
| 전주병 | 1.81146 | 안철수현상 | 1.666928 | 국민의당 | 1.633246 |

국민의당에 긍정적인 영향을 미친 주요 단어를 보면 “공천학살”, “독선적”, “숙의배심원단”, “좌파”, “노무현정신”, “안철수님”, “김종인비례” 등이 보인다. 당시 김종인씨의 셀프공천이 국민의당에는 긍정적인 영향을 미친 것으로 보인다.

아래의 표는 국민의당에 대한 Lasso 회귀 모델링 결과 부정적인 영향을 미친 주요 단어를 정리한 표이다.

표 13. 국민의당 Lasso 결과 부정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|------|----------|-------|----------|------|----------|
| 안빠 | -7.05578 | 나가주세요 | -6.77931 | 이삭줍기 | -6.12092 |
| 민정당 | -5.97931 | 폭력 | -5.75841 | 황창화 | -5.49203 |
| 집결 | -5.26541 | 당적 | -5.07648 | 분열 | -4.85412 |
| 광주경선 | -4.4872 | 홍성문 | -4.17065 | 정계은퇴 | -4.15731 |
| 철수당 | -4.12512 | 우리나라 | -4.09078 | 광주시민 | -3.9453 |
| 불출마 | -3.86713 | 안철수대표 | -3.66567 | 궁물당은 | -3.55903 |
| 철수 | -3.40931 | 새누리당 | -3.29834 | 공관위원 | -3.18861 |
| 예비후보 | -2.96087 | 박그네 | -2.5955 | 기자회견 | -2.48235 |
| 천정배 | -2.4776 | 개누리 | -2.4619 | 민집모 | -2.46031 |
| 선거법 | -2.29321 | 문재인대표 | -2.2375 | 대표자리 | -2.17421 |

국민의당에 부정적인 영향을 미친 주요 단어를 보면 “안빠”, “이삭줍기”, “분열”, “광주경선”, “철수당” 등이 보인다. 당시 안철수 사당화 논란, 더불어민주당 탈당의원들의 국민의당 입당, 야권분열 책임 등이 국민의당에는 부정적인 영향을 미친 것으로 보인다.

아래의 표는 국민의당에 대한 Ridge 회귀 모델링 결과 긍정적인 영향을 미친 주요 단어를 정리한 표이다.

표 14. 국민의당 Ridge 결과 긍정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|--------|----------|-------|----------|--------|----------|
| 비하 | 0.503241 | 정당투표 | 0.44866 | 김종인비례 | 0.447826 |
| 소통 | 0.446094 | 공유 | 0.444749 | 일베 | 0.431697 |
| 세월호 | 0.425822 | 민주정당 | 0.421402 | 당선권 | 0.414872 |
| 낙선대상 | 0.413789 | 비례대표들 | 0.409548 | 더불당 | 0.386611 |
| 안철수의원 | 0.378872 | 민중연합당 | 0.377589 | 이명박정부 | 0.371776 |
| 노무현정신 | 0.364063 | 친문당 | 0.363733 | 친박당 | 0.363731 |
| 남의당 | 0.362418 | 서울시장 | 0.360064 | 지지철회 | 0.356592 |
| 안철수님 | 0.354986 | 예비후보님 | 0.354983 | 이철회 | 0.352852 |
| 리더십 | 0.350829 | 바람 | 0.350682 | 한국정치 | 0.341161 |
| 박영선김종인 | 0.338476 | 개성공단 | 0.338106 | 숙의배심원단 | 0.333994 |

아래의 표는 국민의당에 대한 Ridge 회귀 모델링 결과 부정적인 영향을 미친 주요 단어를 정리한 표이다.

표 15. 국민의당 Ridge 결과 부정적인 영향을 미친 주요 단어

| 단어 | 계수 | 단어 | 계수 | 단어 | 계수 |
|------|----------|---------|----------|--------|----------|
| 운명 | -0.29392 | FTA | -0.24792 | 안철수대표님 | -0.24416 |
| 접근금지 | -0.23599 | 국보위당 | -0.22872 | 노란리본 | -0.22707 |
| 문성근 | -0.22705 | 이정희 | -0.22703 | 간철수 | -0.22428 |
| 피눈물 | -0.2235 | 홍성문 | -0.21769 | 일베충 | -0.21526 |
| 노동자 | -0.2143 | 광주시민 | -0.20907 | 여왕님 | -0.20513 |
| 김무성계 | -0.20507 | 구로 | -0.1976 | 햇불 | -0.19728 |
| 사쿠라들 | -0.19673 | 세월호망치부인 | -0.19365 | 바편애 | -0.19304 |
| 파괴자 | -0.19039 | 목포 | -0.18971 | 후손들 | -0.18717 |
| 곽태원 | -0.18654 | 민생 | -0.18572 | 비례대표공천 | -0.18478 |
| 경주 | -0.18342 | 대한민국주의자 | -0.18238 | 전주병 | -0.18158 |

4.2.4 정당 지지율 예측 결과 및 해석

정당 지지율 예측을 위하여 Lasso 및 Ridge 회귀 분석 알고리즘을 이용한 트위터 데이터 감성분석 결과와 20대 국회의원 선거의 새누리당, 더불어민주당, 국민의당의 정당 지지율을 비교하였다.

20대 국회의원 선거 결과 새누리당, 더불어민주당, 국민의당의 비례대표 선출을 위한 정당투표 결과 새누리당 33.5%, 더불어민주당 25.5%, 국민의당 26.7%로 나타났다.

정당 지지율 예측을 위한 Lasso 회귀 분석을 통한 트위터 데이터 감성 분석의 결과와 극성 정확도를 반영한 각 정당의 지지율은 새누리당 14.7%, 더불어민주당 60.4%, 국민의당 24.7%로 나타났다. 그림 9는 실제 지지율과 Lasso 회귀 분석을 통한 예측 지지율을 비교한 그래프이다.

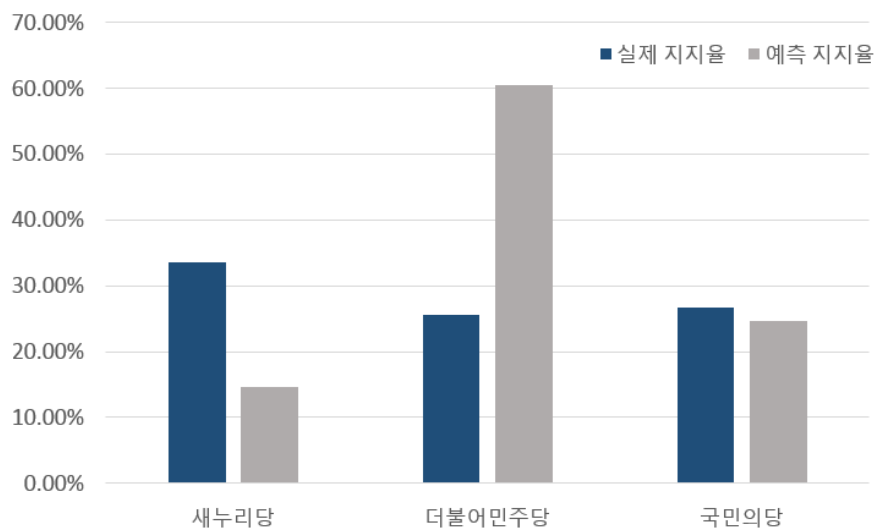


그림 9. 정당 지지율 비교(Lasso 회귀)

실제 정당 지지율과 비교했을 때 새누리당과 더불어민주당의 경우 실제 지지율과 예측 지지율이 다소 차이가 많이 있으며, 국민의당의 경우는 근소한 차이를 있는 것을 볼 수 있다.

정당 지지율 예측을 위한 Ridge 회귀 분석을 통한 트위터 데이터 감성 분석의 결과와 극성 정확도를 반영한 각 정당의 지지율은 새누리당 15.2%, 더불어민주당 57.2%, 국민의당 27.4%로 나타났다. 그림 10은 실제 지지율과 Ridge 회귀 분석을 통한 예측 지지율을 비교한 그래프이다.

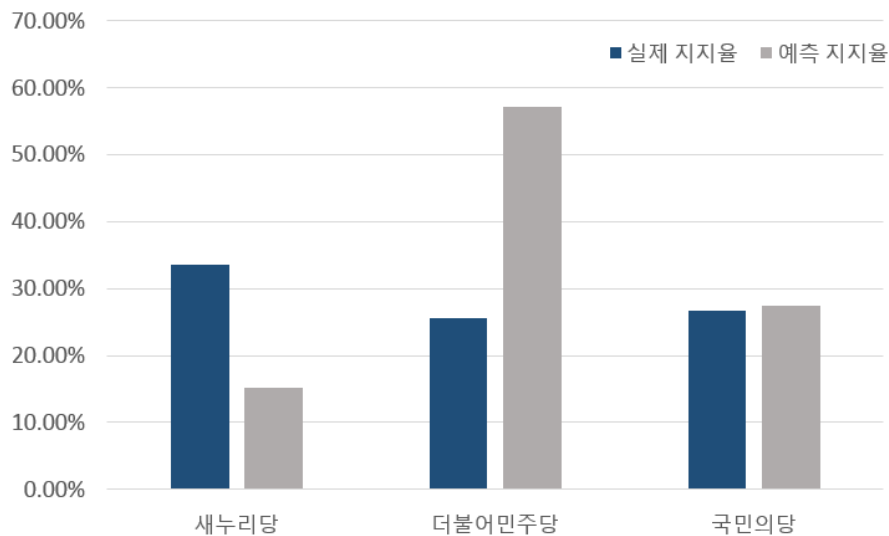


그림 10. 정당 지지율 비교 (Ridge 회귀)

Ridge 회귀 분석을 통한 트위터 데이터 감성분석의 결과 역시 실제 정당 지지율과 비교했을 때 새누리당과 더불어민주당의 경우 실제 지지율과 예측 지지율이 다소 차이가 많이 있으며, 국민의당의 경우는 근소한 차이를 있는 것을 볼 수 있다.

이와 같이 실제 정당 지지율과 Lasso 및 Ridge 회귀 분석을 통한 트위

터 감성분석으로 예측한 지지율이 차이가 나는 것은 소셜 미디어인 트위터가 전 연령대가 사용하기 보다는 주로 20~30대가 주로 사용하며, 우리나라에서 트위터는 개개인의 정치적인 의견을 표현하는 것보다 정당 혹은 정당지지자들이 정당이나 지지하는 후보를 홍보하는 용도로 사용하는 경우가 많고, 지속적인 리트윗(Retweet)을 통해 동일한 트위터 데이터가 다수 생성되는 등의 문제인 것으로 판단된다.

향후 다양한 세대가 트위터를 사용하고 개개인이 자유로운 정치적 의견을 개진하는 성숙한 정치 문화가 정착한다면 이러한 트위터 감성분석이 더욱 효과를 발휘할 수 있을 것으로 판단된다.

제 5 장 결론 및 한계

5.1 연구 결론

본 연구에서는 20대 국회의원 선거 관련 트위터 데이터를 실시간으로 수집하고 Lasso 및 Ridge 회귀 분석 알고리즘을 통해 각 정당에 대한 트위터 데이터의 감성분석 결과 및 예측 정확도 그리고 도출된 결과에 대한 분석하고자 하였고, 실제 정당 지지율과 트위터 데이터 감성분석 결과로 예측된 정당 지지율을 비교 분석하였다. 분석결과를 통하여 얻게 된 결론은 다음과 같다.

첫째, Lasso 및 Ridge 회귀 모델에서 부정 트윗 분류율은 높고 긍정 트윗 분류율이 낮은 것은 긍정 트윗 수가 부정 트윗 수에 비해 상당히 적어 학습데이터로써 긍정 데이터 수가 상대적으로 적으므로 학습을 통해서 긍정은 무시하고 부정의 분류율을 높이려고 하는 Lasso 및 Ridge 회귀 모델의 클래스 불균형 문제가 나타났기 때문에 긍정 트윗 수와 부정 트윗 수의 비율이 비슷해야 하지만 우리나라의 정치문화상 부정 트윗 수가 긍정 트윗 수보다 많음으로 인한 예측의 부정확성이 나타났다. 이를 위해서는 클래스 불균형 문제에 맞는 정교한 모델에 대한 연구가 필요하다.

둘째, 트위터가 제한적인 연령대의 주된 사용과 개개인의 정치적인 의견을 표현하는 것보다 정당이나 지지하는 후보를 홍보하는 용도로 사용하는

경우가 많고, 동일한 트위터 데이터가 다수 생성되는 등의 문제로 인하여 예측의 부정확성이 나타났다. 향후 다양한 세대가 트위터를 사용하고 성숙한 정치 문화가 정착이 필요하고 판단된다.

5.2 연구 한계

본 연구는 20대 국회의원 선거 관련 트위터 데이터의 감성분석을 통하여 정당의 지지율 예측하고 문제점을 고찰하였다. Lasso 및 Ridge 회귀 분석을 통하여 정당 지지율을 예측하였지만 트위터의 특성과 우리나라의 정치 문화 여건 상 현실로 인한 클래스 불균형 문제 등으로 예측이 어려웠다. 이러한 한계에도 불구하고 Lasso 및 Ridge 회귀 분석을 통한 트위터 감성분석 한 것은 의미 있는 성과로 보여지며, 연구를 진행하며 부족한 부분은 향후 연구과제로 남긴다.

첫째, 트위터의 특성과 우리나라 정치문화 여건에 맞는 정교한 트위터 감성분석 모델에 대한 연구가 필요하다고 판단된다.

둘째, 본 연구에서 제안된 방법이 정치 여론의 감성분석 외에 정책 제안이나 회사 제품에 대한 소비자 반응과 같은 감성분석이 요구하는 다양한 다른 영역에서 적용될 수 있기를 기대한다.

참 고 문 헌

- 강만모, 김상락, 박상무. 2011. “빅 데이터의 분석과 활용”, 정보과학회지 30(6) : 25-32
- 이성훈. 2012. “빅데이터 활용 현황”, 한국정보기술학회지, 10(3) : 51-54
- 송영조. 2012. “빅데이터 시대! SNS의 진화와 공공정책”, 한국정보화진흥원
- 금혜성. 2011. “정치인의 SNS 활용”, 한국정당학회보, 10(2) : 189-220
- 한범희. 2014. “텍스트 마이닝을 이용한 서울시장선거 관련 기사와 트윗과의 관계 분석” , 연세대학교.
- 박중호. 2015. “서포트 벡터 머신 학습을 통한 트위터 데이터의 감성분석”, 연세대학교.
- 사공원. 2015 “온라인 감성분석을 이용한 호텔서비스 품질 평가”, 경희대학교
- 차재필. 2012. “빅데이터 시대의 국민공감 선거전략-미 대선사례를 중심으로”, 한국정보화진흥원.
- 이상훈, 최정, 김종우. 2016. “영역별 맞춤형 감성사전 구축을 통한 영화리뷰 감성분석”, 지능정보연구, 22(2) : 97-113.
- 성건, 원규식, 이수원. 2015. “웹 뉴스 댓글 기반 2014 브라질 월드컵 한국 축구 국가대표팀 관련 인물에 대한 감성분석”, 한국스포츠산업경영학회지, 20(2) : 13-28

- 송은지. 2015. “소셜 미디어 상 고객피드백을 위한 감성분석”, 한국정보통신학회논문지, 19(4) : 780-786.
- 윤한중. 2016. “한국어 트위터 데이터의 감성 분석 알고리즘 구현”, 한국과학기술대학교.
- 김동균, 허지용, 조지훈, 박수영, 김용혁. 2011. “기계학습 기반의 감성 트위터 봇”, 한국정보과학회 학술발표논문집, 38(2B) : 379-382.
- 신수정. 2014. “글에서 감정을 읽다 - 감성 분석의 이해”, ITWorld
- Liu, B. 2012. “Sentiment analysis and opinion mining”, *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, Vol. 5, No. 1 : 1-167
- McKinsey & Company. 2011. “Big Data : The Next Frontier for Innovation, Competition, and Productivity”
- M. K. Beyer and D. Laney. 2012. “The Importance of ‘BigData’: A Definition,” Gartner.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer.
- Meier, L., Van De Geer, S., & Bühlmann, P. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1) : 53-71.
- <https://ko.wikipedia.org>

[ABSTRACT]

Investigation of Machine Learning for Predicting the Approval Ratings of Parties in the South Korea's 20th Legislative Election

Ha, Sang Hyun

Major in Industrial Information Management

The Graduate School of Engineering

Yonsei University

Supervised by Prof. Kim, Chang Ouk

As a large amount of data has been created and shared through the social media, big data technology that is able to effectively store and analyze data, has received great attention. Social media is more highlighted than other traditional media because of the strong impact to politics and election as well as information sharing among people.

As the technology of analyzing big data has been increasingly developed and the influence degree of the social media becomes larger, the sentiment analysis that analyzes human emotions, attitude, assessment, and opinions with massive data collected from the social

media in real-time mode is being actively studied.

In this study, a machine learning-based automatic sentiment analysis method for the Twitter data related to the election of the 20th National Assembly is proposed to predict the approval rate of political parties. Two regression models, Lasso and Ridge, were employed to predict the polarity of tweets. Based on the analysis result, we suggest future research area for the machine learning-based sentiment analysis.

I collected 1,048,574 experimental tweets from Twitter through ten keywords - 'election of National Assembly', 'member of National Assembly', 'election', '413 - the date of the 20th election', 'saenuri party', 'the minjoo Party', 'theminjoo', 'aenuri', 'the people's Party', 'official nomination'. The experiment verified that the class imbalance problem of Lasso and ridge regression, environment of politics in Korea, the limited-age Twitter user group, and the promotion of specific candidates via Twitter led to the low performance of the two regression models. From the result, I concluded that sentiment analysis methods pertinent to the characteristic of Twitter and Korean political environment should be further developed.

I expect that this study can be applied to other fields which need the analysis of public opinion such as policy decision-making and the

consumers' responses to new products.

Keyword: Approval rating, Election of National Assembly, Social media, Twitter, Sentiment analysis, Machine learning, Polarity analysis, Lasso regression, Ridge regression