Sentiment analysis 를 이용한 트위터 텍스트 감정분석 모델의 개발

RNN 모델을 사용하여 국내 지방선거 국면에서 트위터 사용자들의 감정으로부터 여론 동향 파악

서다은, 이은후, 이세환
allsilver921@snu.ac.kr, hoospan01@snu.ac.kr, honestlee213@snu.ac.kr

Sentiment Analysis for Public Opinion about Local Election in Twitter

Daeun Seo, Eunhoo Lee, Sehwan Lee

요 약

대표적인 SNS 중 하나인 트위터의 제 8 회 전국동시지방선거 관련 게시글(트윗)을 긍정적, 부정적 어휘에 점수가 매겨진 감정사전을 이용하여 분석하여 정치 정당, 후보자 등에 대한 긍정, 부정 여론을 분석하고 이것을 이용해 RNN 모델을 만들고 학습시킨 후 선거 결과를 예측하였다.

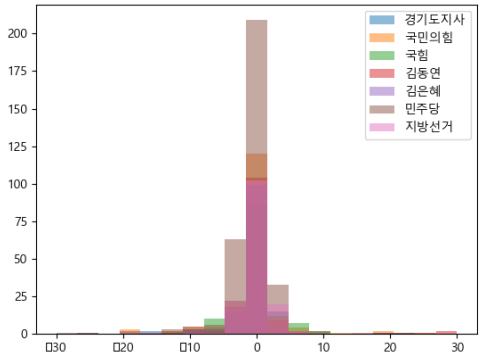
1. 서론

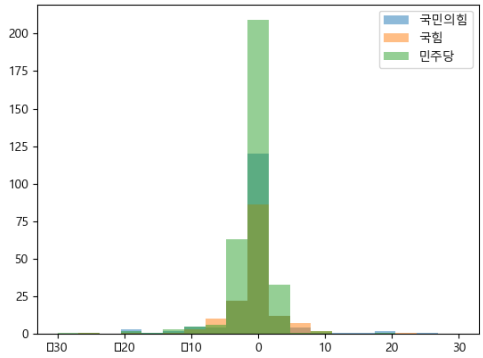
감정분석(Sentiment Analysis)는 텍스트 데이터 가 표상하는 작성자의 의견이나 감성, 평가, 태도 등을 분석하는 일련의 과정을 의미한다. SNS의 활성화로 감정분석이 가능한 영역은 점점 넓어지고 있다. 사람들은 SNS에 자신의 정치적 견해를 올리기도 하는데 이를 감정분석하면 작성자가 어떤 사안 및 인물에 대해 긍정적인 의견을 가지는 지 부정적인 의견을 가지는 지, 그 이유가 무엇인지를 분석해 이것을 정치적 전략을 세우는 데 활용하거나 선거결과를 예측하는데 활용할 수 있을 것이다. 이에 우리는 2022년 6 월 1일에 실시된 제8회 전국동시지방선거 관련 트윗을 분석하고 이것이 실제 선거 결과와 연관이 있는 지 확인하고자 한다.

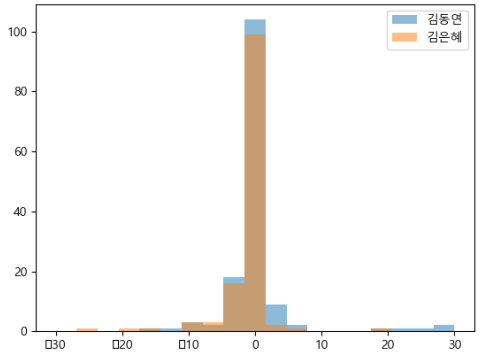
2. 본론

먼저, 트위터의 사전 심사를 거쳐 트위터 수집 권한을 얻어 RNN 모델 학습에 사용할 정치에 관련된 다양한 트윗들과 특히, 지방선거 관련된 트윗들을 수집했다. 사용한 키워드는 "정치", "선거", "지지율", "윤석열", "김건희", "대통령", "청와대", "국회", "장관", "한동훈", "총리", "한덕수", "안철수", "이준석", "박지현", "윤호중", "계파", "당권", "사퇴” 등으로 위의 트위터에 위의 키워드를 검색해서 나온 트윗 중 5 월 26 일부터 5 월 31 일에 작성된 것들을 수집하였다. 특히 이번 지방선거 결과 예측을 위해 이번 지방선거의 주요관심 지역구나 그 지역구에 출마한 인물들 관련 키워드인 "서울시장", "오세훈", "송영길", "부산시장", "계양", "이재명", "윤형선” 이 포함되고 5 월 26 일부터 5 월 31 일에 작성된 트윗들을 추가로 학습에 활용하기 위해 수집하였다. 이렇게 얻은 트윗 중 80%는 학습 데이터로, 나머지 20%는 시험 데이터로 사용하였다. 마지막으로 경기도지사 선거의 예측에 사용할 키워드로 “경기도지사", "국민의힘", "국힘", "김동연", "김은혜", "민주당", "지방선거”를 선택해 5 월 20 일부터 5 월 31 일까지의 트윗을 수집하였다.

KonlPy 라이브러리의 Mecab 을 이용해 트윗의 단어들에 대해 형태소 변환을 시행하였다. 단어에서 실질적 의미를 가지는 실질 형태소들을 남기고 형식 형태소는 제거하는 방식이다. 이렇게 정리한 키워드별 문장들을 KNU 한국어 감성사전을 이용해 단어들에 긍정, 부정 정도에 따라 점수를 매기고 이를 종합하여 문장별로 긍정적인 문장일 경우 1, 부정적인 문장일 경우 0 의 점수를 부여했다.
다음으로는 선행 연구인 “PyTorch 와 TorchText 를 이용한 한국어 감정 분석 연습”을 참고하여 PyTorch 의 내장 인공신경망 모듈과 TorchText 라이브러리를 활용하여 RNN 모델을 만들었다. 모델을 학습데이터로 10 회 학습시킨 후 시험 데이터에 대해 89% 정도의 정확도를 보였다.
이후 6 월 1 일 지방선거 중 경기도지사 선거에 대한 결과 예측을 위해 단어별 긍정, 부정 정도에 따라 1~3, -1~-3 을 부여하고 세부 감정점수를 계산했다.








```
------------- Stitistics -------------
       경기도지사  국민의힘    국힘   김동연   김은혜  민주당  지방선거
count  145.00  145.00  145.00  145.00  129.00  145.00  145.00
mean     2.05   -2.01   -2.41    1.45   -1.16   -0.99   -0.32
std     18.46   15.37   18.43   15.58    4.77    3.86   30.24
min    -54.00 -160.00 -200.00  -72.00  -32.00  -24.00 -330.00
25%      0.00   -1.00   -2.00    0.00    0.00   -1.00    0.00
50%      0.00    0.00    0.00    0.00    0.00    0.00    0.00
75%      0.00    0.00    0.00    0.00    0.00    0.00    0.00
max    148.00   24.00   36.00  150.00   20.00   10.00   92.00
```

```
------------- Stitistics -------------
경기도지사
equal 0: 59.3
-5 under: 6.2%    5 over: 5.5%
-10 under: 4.1%    10 over: 4.1%


국민의힘
equal 0: 51.0
-5 under: 11.0%    5 over: 6.9%
-10 under: 6.2%    10 over: 3.4%


국힘
equal 0: 51.7
-5 under: 13.1%    5 over: 7.6%
-10 under: 4.8%    10 over: 2.8%


김동연
equal 0: 62.8
-5 under: 5.5%    5 over: 6.9%
-10 under: 2.1%    10 over: 5.5%


김은혜
equal 0: 62.8
-5 under: 6.9%    5 over: 1.4%
-10 under: 3.4%    10 over: 0.7%


민주당
equal 0: 49.0
-5 under: 7.6%    5 over: 0.7%
-10 under: 3.4%    10 over: 0.7%


지방선거
equal 0: 54.5
-5 under: 3.4%    5 over: 5.5%
-10 under: 2.8%    10 over: 4.1%
```

3. 결론

김동연 후보의 평균적인 점수가 김은혜 후보보다 높게 나타났다. 이는 실제 경기도지사로 김동연 후보가 당선된 사실과 일치한다. 또한, 김동연 후보에 대한 문장의 점수가 +5, +10 이상인 비율이 김은혜 후보보다 월등히 높은 것으로 나타나 김동연 후보는 열성적인 지지층을 보유한 것으로 보인다. 정당에 대한 평균적인 점수도 국민의 힘보다 민주당이 높았던 것으로 나타났다.
많은 여론조사에서 김은혜 후보의 우세를 점쳤으나 트위터를 이용한 데이터 분석으로는 실제 결과와 같이 김동연 후보의 승리를 예측할 수 있었다. 그러나 이러한 트위터를

이용한 분석은 한계점을 가지는데 첫째로 얼마나 많은 사람들이 자주 트위터를 사용하는 지의 문제, 즉 유의미한 사용량을 가지는 지의 문제가 있다. 둘째로는 사용자들이 정치적으로 편향되어 있지는 않은 지에 대한 연구가 수반되면 더 정확한 분석이 가능할 것이다.

4. 참고문헌

- PyTorch와 TorchText를 이용한 한국어 감정 분석 연습, https://github.com/lih0905/korean-pytorch-sentiment-analysis
- KNU 한국어 감성사전, https://github.com/park1200656/KnuSentiLex