



# 2024 미국 대선 데이터 분석 및 예측



통계데이터과학전공  
2019E7022 김부겸

# 목차

1

주제 선정 이유

2

데이터 수집

3

데이터 전처리

4

데이터 시각화

5

모델링 및 예측

6

결론

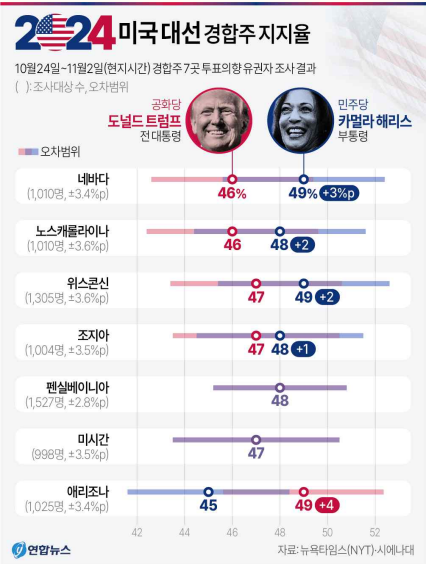
# 주제 선정 이유



해리스 vs 트럼프

2024년 11월 5일  
미국 대통령 선거 실시

# 주제 선정 이유



홈 > 정치 > 국회·정당·정책

## [데이터로 본 정치민심] 與당대표 선거 D-1, 검색량으로 보면 누가될지 안다

입력 2021.05.01 13:59:07 수정 2021.06.04 16:54:49



### ■[네이버트렌드-썸트렌드]

홍영표, 키워드에서도 확연한 '친문' 성향  
우원식, 검색량 저조하지만 '민생'과 '민평련'  
송영길, '이재명'·'러시아 백신'으로 이슈 선점

더불어민주당 당대표와 최고위원을 뽑는 5·2 전당 대회가 하루 앞으로 다가왔다. 이번엔 뽑힐 당 지도부는 4·7 재보궐 선거 참패로 어수선했던 당을 쇄신하고 정권재창출을 목표로 당을 이끌어야 한다. 내년 3월 대통령 선거까지 누가 운전대를 잡을 지 관심이 집중되는 이유다.

## 여론 조사

투표의향 유권자 조사 결과  
지지율 경합 지역 다수

## 빈도수

검색량(빈도수)에 따라  
각 후보자별 지지율 예측 가능

## 감성분석

대중들의 반응을 이용  
(Python 이용)

# 데이터 수집



대선 토론



뉴스 기사



X(Twitter)

대선 관련 텍스트 데이터를 수집하여 2024 미국 대선 예측  
(수집 기간 : 2024.09.01 ~ 2024.10.31)

# 데이터 수집

- 대선 토론



# kaggle

2024 9월 10일 토론 데이터 수집

- 9월 5일과 10월의 토론은 무산됨


# 데이터 수집

- 뉴스 기사



 CBS

 npr online news only

 reason



# 데이터 수집

- X(twitter)





# 데이터 전처리



## 데이터 수집

각 매체 별 데이터 수집

## 토큰화

NLTK의 word\_tokenizer를 사용

## Pos tag + Lemmatizing

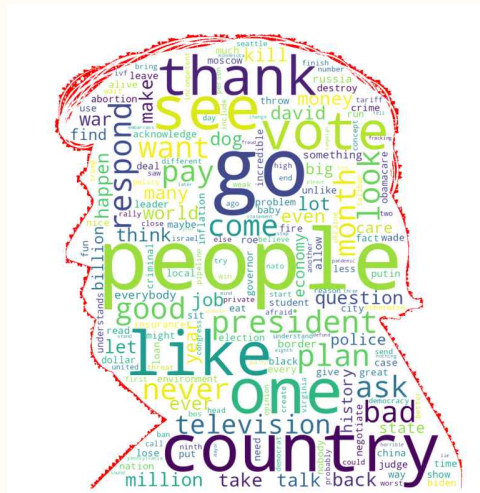
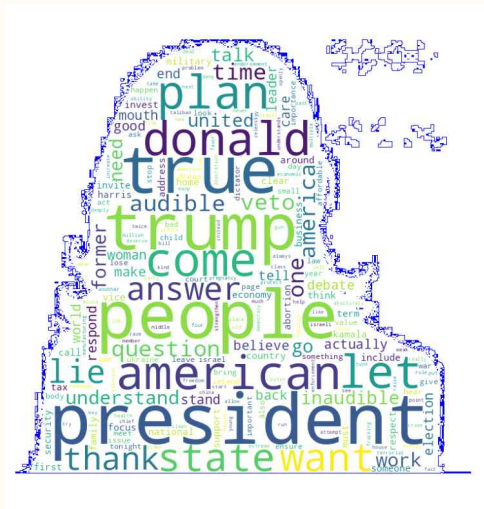
정확한 추출을 위해 토큰별 품사를 태그해 어간 추출 진행

## 역토큰화

정제된 토큰을 역토큰화를 통해 문장단위로 통합

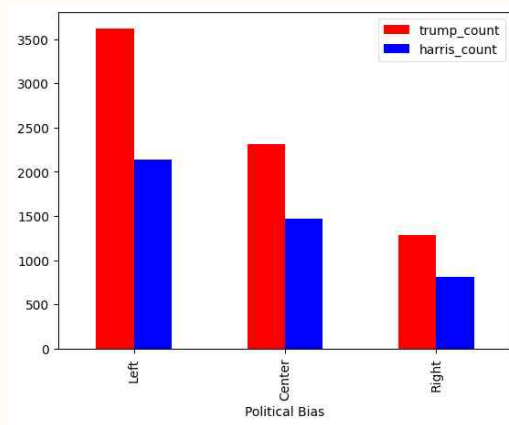
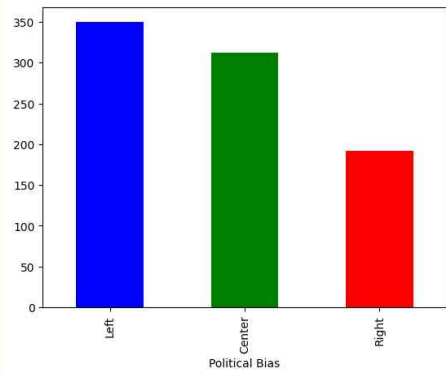
# 데이터 시각화

## - 대선 토론



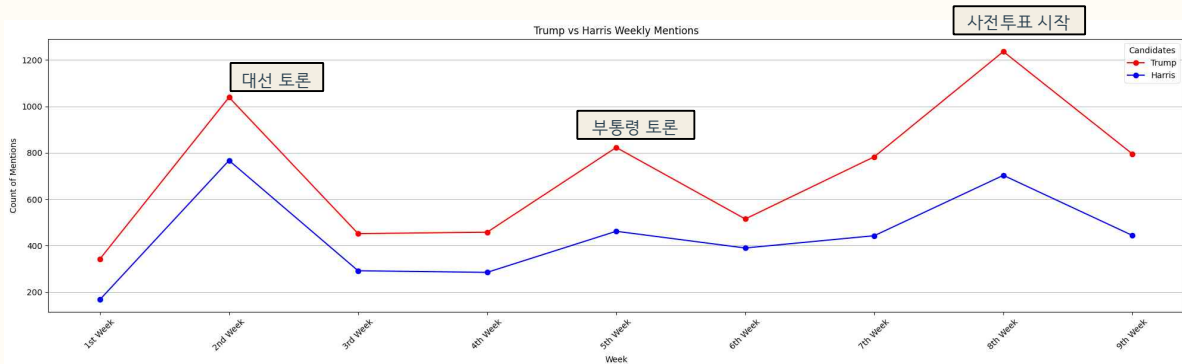
# 데이터 시각화

- 뉴스 기사



# 데이터 시각화

- 뉴스 기사

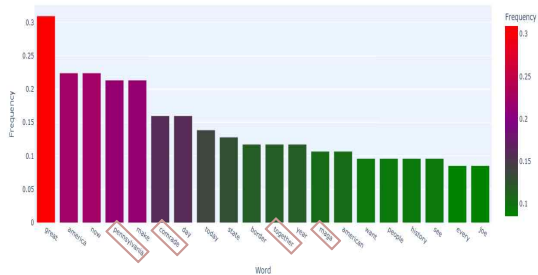


각 주별 후보자 언급 횟수

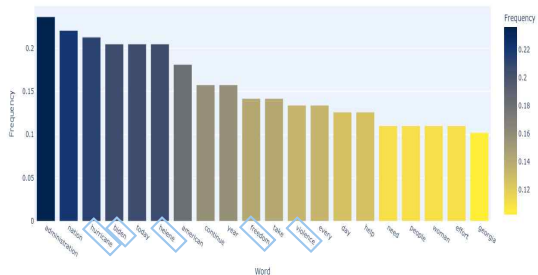
# 데이터 시각화

- X(twitter)

Trump's Twitter by TF-IDF



Harris's Twitter by TF-IDF



후보자가 작성한 트위터 키워드 분석

- $X(\text{twitter})$



후보자에게 보낸 트위터 키워드 분석

# 데이터 시각화

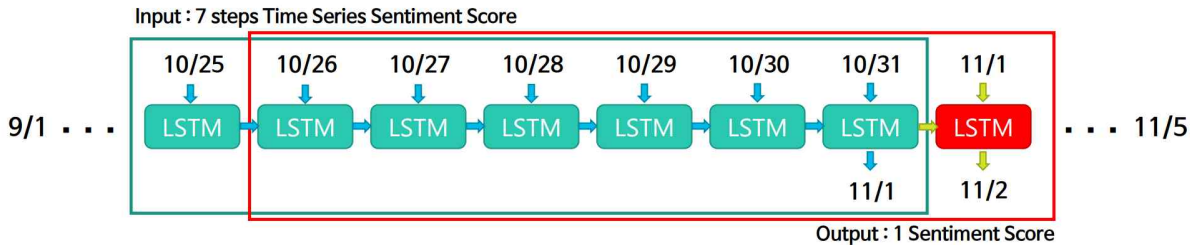
- X(twitter)



감성구분	점수
긍정	compound score $\geq 0.05$
중립	$-0.05 < \text{compound score} < 0.05$
부정	compound score $\leq -0.05$

# 모델링 및 예측

- LSTM

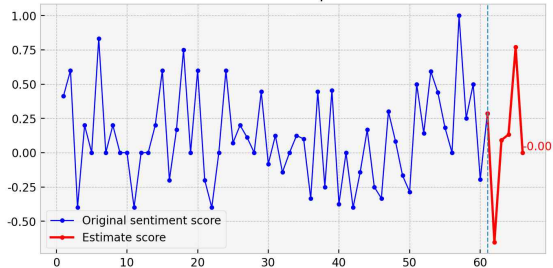




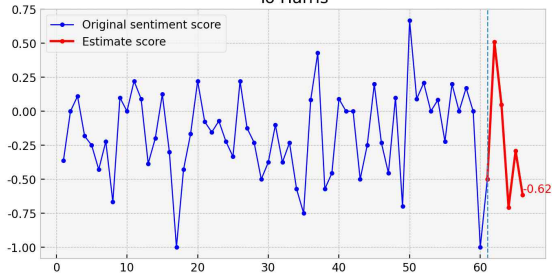
# 모델링 및 예측

- LSTM

To Trump



To Harris



후보자에게 보낸 트위터 감성 분석 예측



# 결론

## 빈도수

뉴스에서 트럼프가 해리스보다  
많이 언급됨

## 감성 예측

LSTM 분석 결과  
트럼프의 감성점수가  
해리스 보다 높게 나옴

## 한계

모든 매체와 모든 사람을  
조사한 것이 아니기에  
정확하지 않을 수 있음

감사합니다