KISDI STAT REPORT

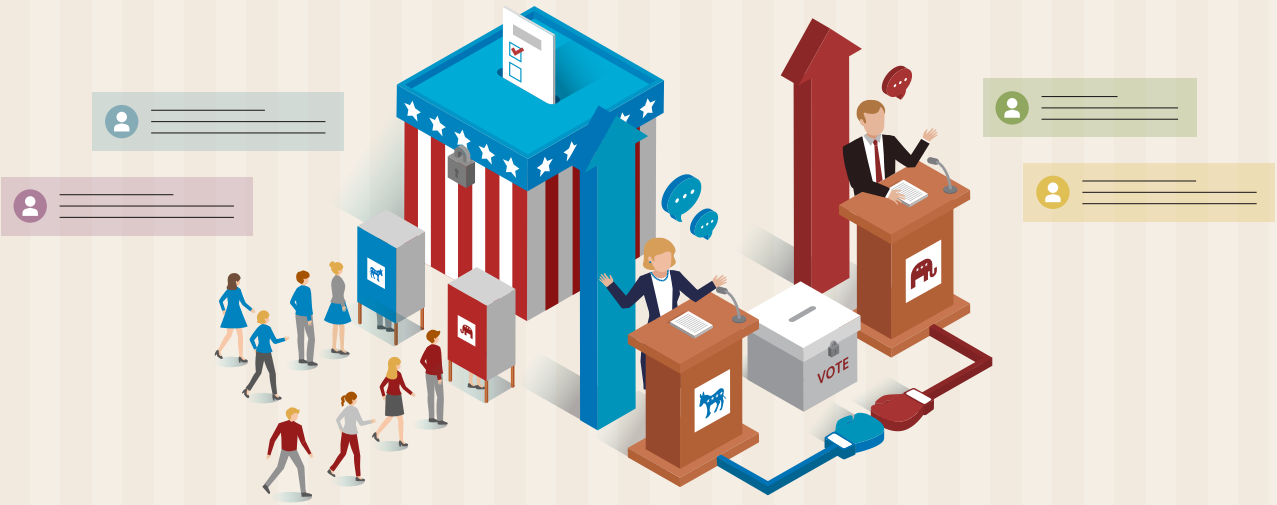
KISDI STAT Report는 ICT, 방송, 통신의 최신 이슈를 데이터에 기반하여 분석하는 온라인 정기간행물입니다. KISDI 연구자 및 산학연 전문가들이 참여하여 매월 30일에 온라인으로 배포합니다.

ISSN 2384-1672

정보통신정책연구원 KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

Youtube 댓글을 통해 보는 미국 대선 지지 정당 감성분석

Analysis of Political Sentiment Towards U.S. Presidential Candidates Based on YouTube Comments

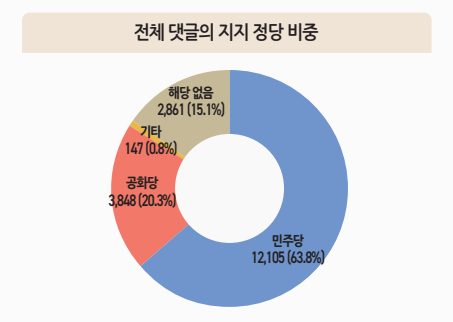


노희용 (연구위원, ICT통계정보연구실)

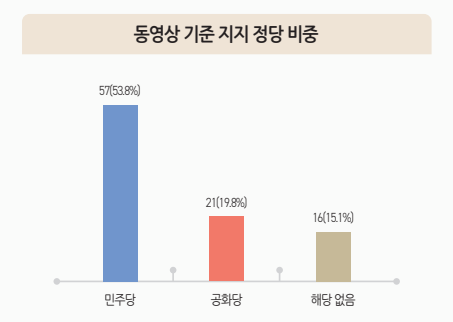
KEYWORDS 미국 대선, 거대언어모형, 감성분석, 라마
US presidential election, large language model, Llama, sentiment analysis

미국이 세계 무대에서 미치는 사회경제적, 정치적 영향력의 높은 중요성 때문에, 최근 많은 나라와 국제기구에서는 약 두 달 남은 2024년 미국 대선의 향방에 많은 관심을 보인다. 차기 미국 정부의 정책 기조를 분석하고 이에 따른 대응 방안 등을 모색하는 노력이 이루어지고 있는 가운데, 본 보고서 또한 이와 같은 노력의 일환으로 미국 대선의 향방을 유튜브 댓글을 활용해 분석했다. 분석 과정은 크게 세 가지 단계로 진행되었다. 먼저, 유튜브 데이터 API를 사용하여 가장 많이 조회된 유튜브 비디오를 식별하고, 이들 비디오의 모든 댓글을 수집한다. 다음으로는 메타에서 공개한 Llama3.1을 사용해 각 댓글에서 언급된 지지 정당을 분류한다. 마지막으로, 감성 분석으로 활용되는 VADER 알고리즘으로 모든 댓글의 감성 점수를 파악한다. 이를 통해 산출된 결과를 보면 민주당이 더 긍정적인 관심을 받는 것으로 나타난다.

The 2024 Unites States(U.S.) presidential election is now about two months away. Given the significance socioeconomic and political influence of the U.S. on the world stage, all eyes are on the outcome of this election. As a way to gauge the direction of the U.S. presidential race, this report analyzes the Youtube comments. Three steps are undertaken: (i) First, the most-viewed Youtube videos are identified using the Youtube Data API, and all comments from these videos are scrapped. (ii) Next, Meta's Llama3.1 is adopted to classify the supporting parties mentioned in each comment. (iii) Finally, the VADER algorithm is used to capture the sentiment scores of all the comments. The analysis reveals that the Democratic Party is receiving more positive attention.



03 분석 결과 ▶▶4



03 분석 결과 ▶▶4

Aug. 30, 2024

24/15

24-14호 (24.07.30)
AI 챗봇 인지도 및 이용행태
오윤석

24-13호 (24.07.15)
고용노동통계로 살펴본 산업별 고용 특징 비교
서영선

24-12호 (24.06.30)
디지털 전환(Digital Transformation) 체감도 및 영향력 인식
한은영

Youtube 댓글을 통해 보는 미국 대선 지지 정당 감성분석

Analysis of Political Sentiment Towards U.S. Presidential Candidates Based on YouTube Comments

노희용 | 연구위원, ICT통계정보연구실

01 배경

- 2024년 대선까지 약 2개월을 앞둔 미국의 정세 변화에 대한 세계의 관심이 매우 큰 상황
- 이는 어떤 당에서 국가수반의 자리를 차지하느냐에 따라 미국이 전 세계에 미치는 강대한 영향력의 방향이 결정되기 때문임
- 예를 들어, 사실상의 차기 대권 경쟁 관계로 여겨지는 민주당(Democratic party)과 공화당(Republican party)의 경제정책 기조를 살펴보면*, 對중국 정책을 빼고 전략적 차별화가 두드러짐

  * 한국경제인협회 보도자료 참고(2024.08.14)

- 특히 민주당과 공화당의 정보통신기술(ICT, Information and Communications Technology) 정책을 보면*, 민주당은 반독점법 집행 강화와 빅테크 기업의 분할을 검토하는 반면 공화당은 대기업의 자유로운 경영 활동을 지원하고 인수합병 규제를 완화하고자 하는 등 서로 완전히 반대되는 방향을 관찰할 수 있음

  * 정보통신산업진흥원 8월 월간 동향 참고(2024.08.30.)

- 이처럼 상반된 정책 방향에 따라 전세계에서는 미국 대선 결과에 따른 장기 전략 및 정책 방향에 대한 논의가 계속해서 거듭되고 있음
- 동시에 대선 결과를 점쳐보고자 하는 노력 또한 이뤄지고 있는데, 본 고 또한 대표적인 동영상 플랫폼인 Youtube를 바탕으로 2024년 미국 대선에 대한 여론을 피상적이나마 실험적인 방법으로 탐색해보고자 함

  * 여기서 '피상적이나마 실험적인 방법'이란 논리적인 이론과 깊은 고민을 바탕으로 하는 연구(research)가 아닌 단순하고 직관적인 일회성 시도로 방법론적 여흥에 가까움

02 분석 방법

1. 활용 데이터

- 본 고에서는 Youtube 플랫폼에서 한국시간(GMT +9) 기준 2024년 8월 21일 00:00 AM에 "USA election 2024"를 쿼리*로 할 때, 가장 조회수가 많은 동영상의 댓글을 분석 데이터로 함

  * 앞서 설명했지만 모든 분석은 본 고의 작성을 위해서만 새롭게 진행된 까닭에 동영상 검색 역시 단순한 한 개의 쿼리만을 활용함

- 동영상 목록을 불러오기 위해 Google for Developers의 Youtube Data API를 활용했고, API의 사용량 제한이 있어 댓글 수집*은 웹 스크랩을 통해 직접 수행함

  * 대댓글(reply)은 소수의 사용자(user) 간 대화의 형식으로 진행되는 경우가 잦아 제외하고, 댓글(comment)만 2024년 8월 21일 00:00 AM부터 순차적으로 수집함

- Youtube Data API의 검색은 일반적인 순차적 페이지 뷰(page view)로 이루어지지 않고, 검색 지점으로부터 일정 반경(radius) 내에서 가장 접근이 잦은 동영상만 페이지 토큰(page token)을 기반*으로 이루어짐

  * 일반적으로는 페이지 번호(page number)에 따라 목록(list)화된 개체가 관찰되지만, Youtube Data API에서는 일련의 문자 나열로 이루어진 페이지 토큰(page token)별로 제한된 수의 동영상만 확인할 수 있음

  * 이는 Youtube 플랫폼이 전 세계적으로 활용되고 있기에 단일 서버에서 데이터 출납(retrieval)이 이루어지지 않고, 실시간 변화가 계속해서 이루어지는 동적 구조를 다루기 위함과 동시에 API 활용에 제한을 두기 위한 것으로 판단됨

- 따라서 본 고에서는 검색 지역을 미국으로 설정하고 영어로 작성된 비디오 중 Youtube 플랫폼 내에서 '신뢰할 수 있는 채널'만 고려했기 때문에, 페이지 토큰이 끝날 때까지* 총 106개의 동영상 주소를 확보할 수 있었고, 수집 당시 동영상 전체에 대해 18,961개의 댓글을 수집함

  * 관련 동영상을 이어서 알려주지 않을 때까지

2. 데이터 가공

- 특별한 데이터 가공은 거치지 않고 UTF-8 인코딩 기준 깨지지 않는 댓글의 텍스트만 활용함

3. 데이터 분석 방법

- 크게 두 가지 단계를 거치는데 첫 번째는 댓글이 지지하는 정당을 판별하고, 두 번째는 댓글의 내용에 대한 감성을 확인함

- 본 고에서는 대부분이 영문으로 작성된 댓글이 지지하는 정당을 판별하기 위해 Meta에서 공개한 거대언어모형(LLM, Large Language Model)인 Llama3.1*을 활용함

  * 파라미터 크기가 각각 80억(8b), 700억(70b), 4,050억(405b)개인 3개의 모형이 있는데, 본 고에서는 가장 작은 8b 모형을 활용함

- 판별을 위한 프롬프트를 던지기 위해서 하나의 system 규칙 지정, 두 개의 assistant를 활용한 부가 정보 제공, 한 개의 user 창을 활용함

  * system은 LLM이 총 10개의 답변 옵션(민주당, 공화당, 녹색당, 자유당, 헌법당, 미국연대당, 사회주의해방당, 금주당, 무소속, 해당하지 않음) 중 하나만 골라 단답형으로 줄 수 있도록 규칙을 설정함

  * 첫 번째 assistant에서는 총 9개 당에 대한 대권 주자와 부통령 후보에 대한 정보를. 두 번째 assistant에서는 각 당의 당색과 슬로건에 대한 정보를 제공함

  * user 프롬프트에서는 비디오의 제목과 댓글 텍스트를 주고, 해당 댓글이 어떤 당을 지지하는지를 추론할 수 있도록 질문을 함

- 각 댓글의 분석은 소셜 미디어의 텍스트에 대한 감성분석을 위해 개발된 사전(lexicon) 기반의 VADER(Valence Aware Dictionary and sEntiment Reasoner) 알고리즘*을 활용함

  * Hutto & Gilbert (2014)

  * VADER는 규칙 기반으로 사전에 수록된 단어의 긍정과 부정의 극성 값에 따라 전체 문장의 감성을 수치화하여 알려주는 빠르고 강력한 방법으로, utf-8 인코딩 환경에서의 이모티콘에도 극성 값을 부여하고 있어 오피니언 마이닝에서 널리 활용됨

03 분석 결과

1. 지지 정당 판별 결과

- Llama의 특성상 기본적인 답변은 대체로 왜 이런 답변을 내놓았는지에 관한 생각(또는 추론 과정)과 질문에 대한 답을 문장 혹은 문단 단위로 생성하며, 특히 질문에 대한 답을 모형이 내리기 힘들 때 이와 같은 추론을 길게 설명하는 경향이 있음

- Youtube 플랫폼의 특성상, 그리고 정책 인사에 대한 주관적인 판단이 들어가는 특성상, 대권 주자 관련 영상의 댓글은 (1) 줄임말* 또는 별명 사용, (2) 의미 없는 댓글, (3) 영상에 등장하는 다른 인물에 대한 의견, (4) 정책적 비판이 아닌 인물 또는 유사 캐릭터 등을 활용한 원색적인 인물 비난 등으로 지지 정당 판단이 쉽지 않은 경우가 많은데, Llama3.1는 줄임말과 당색, 슬로건 등의 줄임말은 잘 파악하는 반면 의미 없는 댓글, 영상에 등장하는 다른 인물에 관한 의견은 판단을 잘 내리지 못하는 경우가 잦았음

  * 예를 들면, VP(Vice President), MAGA(Make Great America Again), JD(James David), Female Obama(Kamala Harris를 의미), GOP(Grand Old Party) 등이 이에 해당함

- 본 고에서는 전체 18,961개의 댓글 중 모형의 판단 혼동으로 설명이 길어지는 경우인 741건은 전수 직접 수정했고, 그 외 줄임말 등의 확인을 위해 전반적으로 휴리스틱 점검을 수행함

- 그 결과, 비디오 내용과는 관계없이 각각의 지지 정당에 대한 의견 비율은 [그림 1]과 같이 민주당 지지 비중이 압도적으로 높았으며 그 뒤를 공화당이 추격하고 있음

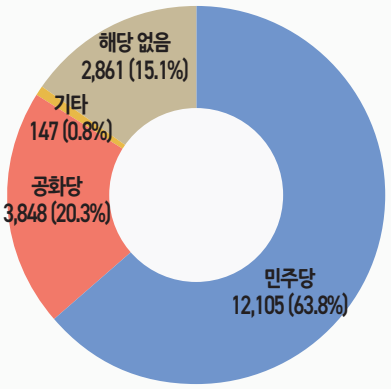
  * 기타 147건의 경우 자유당 지지 댓글 127개, 녹색당 지지 댓글 17개, 사회주의 해방당 지지 댓글 3건으로 구성되며, 비중이 작아 기타로 한데 묶음

  * '해당 없음'으로 분류된 2,861개의 댓글은 다른 나라 혹은 개인의 사연, 특별한 텍스트 없이 동영상 시간 링크만 달아두는 등의 의미 없는 댓글을 의미함

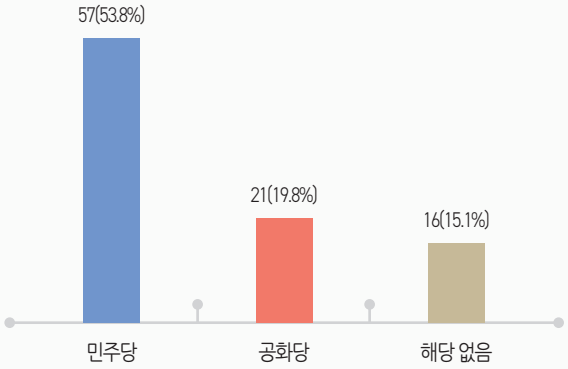
- 다만, 이와 같은 결과는 특정 정당에 관한 내용을 다룬 비디오의 댓글 수가 많은 경우 편향 가능성이 존재하기 때문에, 특정 정당이 우세한 비디오의 수로 고려하면 [그림 2]와 같음

  * '해당 없음'은 정당과 해당 없는 댓글의 비중이 전체 댓글 대비 50% 이상이거나, 양강 정당인 공화당과 민주당 각각의 지지 댓글 비중이 같은 경우로 정의함

[그림 1] 전체 댓글의 지지 정당 비중 (단위: 건)

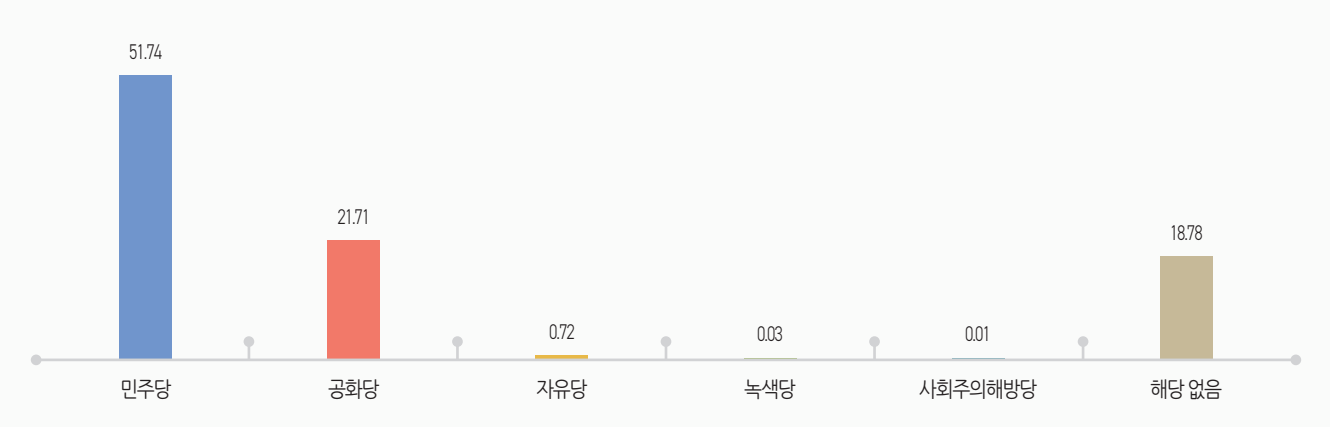


[그림 2] 동영상 기준 지지 정당 비중 (단위: 개)



- [그림 3]은 댓글 수와 관계없이, 각 정당을 지지하는 비중의 합을 비교한 것으로 [그림 2]의 결과와 매우 유사한 형상을 보여줌

[그림 3] 동영상 별 지지 정당 댓글 비율의 합 (단위: 점)



주: 지지 정당 댓글 비율의 합이란 각 동영상별 특정 정당을 지지하는 댓글의 비율(따라서 각 동영상 별 모든 지지 정당 댓글 비율의 합은 1)을 모든 영상에 대하여 합한 것으로, 예를 들어 동영상 3개가 있고 각각 민주당을 지지하는 댓글의 비율이 0.2, 0.4, 0.5라면 민주당 지지 댓글 비율의 합은 1.1이 됨

2. 감성분석 결과

- 감성분석은 결국 각각의 지지 정당에 대한 감성을 확인하려는 방법으로, 포괄적인 감성 분포를 확인하면 [그림 4]와 같음

[그림 4] 지지 정당별 댓글 긍부정 점수 합 분포 (단위: 점)

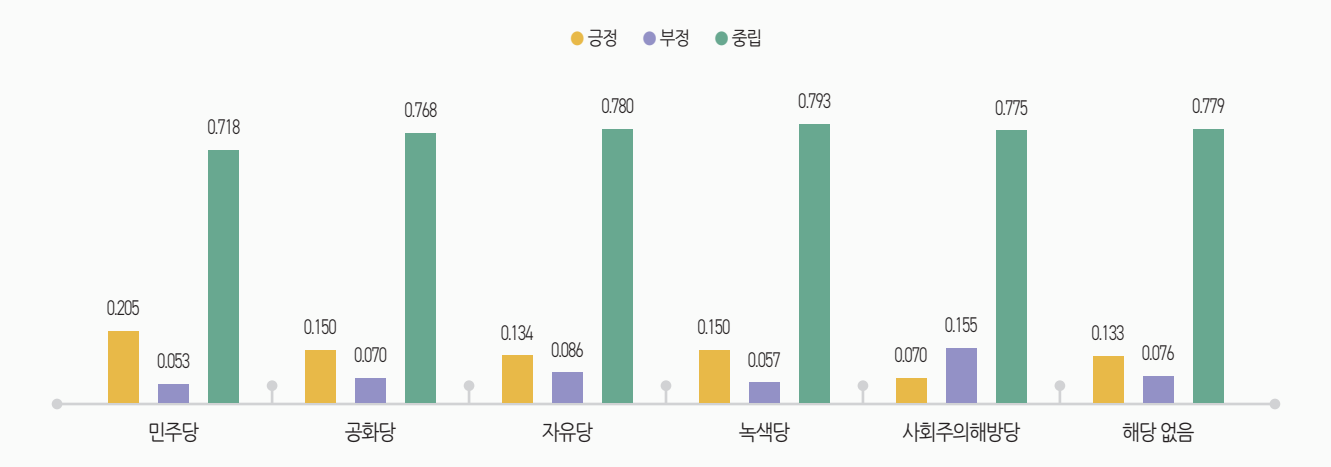
주: 지지 정당별 댓글 긍부정 점수의 합이란 LLM을 통해 댓글별로 지지 정당이 판별된 후, VADER 알고리즘의 결과로 댓글별 산출되는 긍정, 부정, 중립의 감성 점수가 있을 때, 모든 영상의 특정 정당 지지 댓글 감성 점수를 더한 것을 의미함

- 지지 정당과 관계없이 중립 점수가 높은 가운데, 민주당 지지 댓글의 긍정적 감정 대비 부정적 감정의 비율(약 0.26)은 공화당 지지 댓글의 긍정적 감정 대비 부정적 감정의 비율(약 0.47)에 비해 낮음*을 확인할 수 있음

  * 이는 공화당 지지층의 외연 확대 가능성이 크지 않은 가운데, 민주당 지지층에 대한 약화를 꾀하는 비난 전략이 Youtube 플랫폼에서도 작동하고 있을 가능성을 시사함

- [그림 5]는 각각의 지지 정당별 전체 감성 점수의 합을 댓글 수로 나눈 분포로 이 또한 민주당에 대한 지지 댓글의 긍정적 감정 평균이 공화당의 긍정 감정 평균보다 높고, 민주당 지지 댓글의 부정적 감정 평균이 공화당의 부정 감정 평균보다 낮음을 확인할 수 있음

[그림 5] 지지 정당별 댓글의 평균 긍부정 점수 분포 (단위: 점)



04 결론 및 방법론적 한계점

- 분석 결과 8월 21일을 기준으로 Youtube 플랫폼에서는 민주당의 Kamala Harris 후보가 우세한 것으로 판단되며, 이는 단순히 민주당을 지지하는 댓글의 수나 동영상의 수뿐만 아니라 감성 분석에서도 긍정적인 의견이 다른 경쟁 정당에 비해 높다는 점에서 확인할 수 있음

- 이와 같은 결과는 [그림 6]과 같이 8월 30일 발표된 미국 여론조사 결과에서도 확인할 수 있는데, 심지어 한 달전(7월) 민주당이 열세로 조사된 The Wall Street Journal에서도 우위를 점하고 있는 것으로 나타남

[그림 6] 미국 대선 관련 여론조사 결과(2024.09.03. 확인)

- 본 고의 분석 결과는 흥미로운 결과를 제공하지만, (1) Youtube의 특성상 수집 데이터의 대표성과 댓글을 바탕으로 하는 여론 확인의 대표성 여부를 논리적으로 검토할 수 없다는 점, (2) 적극적인 댓글 사용자의 지지 정당 편향 가능성, (3) 댓글 생성 봇(bot) 가능성, (4) Llama3.1 모형의 댓글 지지 정당 판별 정확도 확인이 어려운 점 등에서 방법론적 한계를 가지고 있음
- 특히 8월 21일을 기준으로 수집된 snapshot 데이터 기반의 자료라는 점에서 언제든지 시점에 따라 다른 결과가 나타날 수 있기에 재현이 어렵다는 점에서 교차 검증이 쉽지 않음
- 다만, Youtube 플랫폼이라는 소셜 미디어의 관점에서 지속적인 자료 수집 및 분석 모니터링을 수행한다면 정책적 또는 산업적 의사결정의 참고자료로 활용할 수 있다는 점에서 본 고의 가치가 있을 것으로 판단됨

05 참고자료

한국경제인협회(2024.8), “2024 美 대선 양당 경제정책 기조 분석”, https://www.fki.or.kr/main/news/statement_detail.do?bbs_id=00035718&category=ST (2024년 8월 21일 접속)

정보통신산업진흥원(2024.8), “미국 대선 결과에 따른 ICT 정책 변화 전망”, https://www.globalict.kr/news/trend/weekly.do?menuCode=010200&knwldNo=144272 (2024년 9월 1일 접속)

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

네이버뉴스검색결과(2024.9),https://search.naver.com/search.naver?where=news&query=%EB%AF%B8%EA%B5%AD%20%EB%8C%80%EC%84%A0%20%EC%84%A4%EB%AC%B8%EC%A1%B0%EC%82%AC&sm=tab_opt&sort=0&photo=0&field=0&pd=0&ds=&de=&docid=0520002080856&related=1&mynews=0&office_type=0&office_section_code=0&news_office_checked=&nso=so%3Ar%2Cp%3Aall&is_sug_officeid=0&office_category=0&service_area=0 (2024년 9월 3일 접속)

KISDI STAT REPORT

과월호 목록

정보통신정책연구원 KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

"지난 호는 KISDI STAT 홈페이지 http://stat.kisdi.re.kr 에서 보실 수 있습니다"

발간번호	제목	저자	발간일
24-14호	AI 챗봇 인지도 및 이용행태	오윤석	2024-07-30
24-13호	고용노동통계로 살펴본 산업별 고용 특징 비교	서영선	2024-07-15
24-12호	디지털 전환(Digital Transformation) 체감도 및 영향력 인식	한은영	2024-06-30
24-11호	생산연령인구의 4차 산업혁명과 디지털 전환 인식 비교	하승희	2024-06-15
24-10호	유료방송 가입자의 미디어 소비와 OTT	정용찬	2024-05-30
24-09호	세대별 SNS 이용 현황	김윤화	2024-05-15
24-08호	뉴스 데이터를 활용한 사회갈등의 탐색	서영선	2024-04-30
24-07호	휴대용 전자기기 브랜드 선택에 관한 탐구	장신재	2024-04-15
24-06호	아동·청소년의 미디어 이용행태와 미디어 이용 제한	김윤화	2024-03-30
24-05호	웨어러블 기기 이용현황 분석	박지원	2024-03-15
24-04호	SNS 이용과 디지털콘텐츠 이용 간 관계 분석	장현지	2024-02-28
24-03호	인스턴트 메신저 서비스와 동영상 콘텐츠 제공 서비스 이용자의 미디어 이용행태	하승희	2024-02-15
24-02호	장르별 방송프로그램 품질과 시청자 인식	한은영	2024-01-30
24-01호	스마트폰과 TV의 시간 점유율 경쟁	정용찬	2024-01-15
23-24호	2023년 방송산업 실태조사 주요 결과	한은영, 오윤석, 하승희	2023-12-30
23-23호	2023년 한국미디어패널조사 주요 결과	정용찬, 김윤화, 오윤석	2023-12-15
23-22호	미디어 리터러시 역량의 지역 격차	고흥석	2023-11-30
23-21호	누가 온라인 정보를 생산하고 확산시키는가? 비판적 사고를 중심으로	오주현	2023-11-15
23-20호	코로나19 전후 개인 미디어 이용 행태 및 감정 상태	장신재	2023-10-30
23-19호	여가용 디지털 콘텐츠 지출현황 분석	장현지	2023-10-15
23-18호	뉴스 데이터 기반 지수들의 특징 비교	서영선	2023-09-30
23-17호	TV시청자의 시청 경험과 인식 변화	한은영	2023-09-15
23-16호	태블릿PC 이용자의 미디어 이용행태 분석	오윤석	2023-08-30
23-15호	ICT 고용지수의 의미와 시사점	노희용	2023-08-15
23-14호	방송사업자 광고매출 현황 분석	박지원	2023-07-30
23-13호	모바일 간편 송금 및 결제 서비스 이용자의 미디어 이용 행태	하승희	2023-07-15
23-12호	랜선 공연의 등장과 문화 예술 향유 추이 분석	김나연	2023-06-30
23-11호	MZ세대의 미디어이용 특징	김윤화	2023-06-15
23-10호	1인 가구의 확산과 미디어 이용 변화	정용찬	2023-05-30
23-09호	주요국 OTT(비디오 스트리밍) 서비스 이용현황 분석	한은영	2023-05-15
23-08호	메타버스 이용 현황 및 이용자 특성	김윤화	2023-04-30
23-07호	ICT이머징이슈발굴 시스템을 활용한 2023년 1분기 ICT 산업 동향 분석	노희용	2023-04-15
23-06호	OTT 서비스 플랫폼별 이용행태 비교	오윤석	2023-03-30
23-05호	워크넷으로 살펴본 ICT 산업의 구인과 취업	서영선	2023-03-15
23-04호	해외 오디오 스트리밍 서비스 이용현황 분석	박지원	2023-02-28
23-03호	SNS 이용시간이 삶의 만족도와 자아존중감에 미치는 영향	김나연	2023-02-15
23-02호	코로나와 미디어 이용	정용찬	2023-01-30

KISDI ICT통계정보연구실

KISDI에서는 다음과 같은 정부 승인통계 조사를 매년 진행하고 있습니다. 조사 결과는 정부 정책 수립과 민간 기업의 경영계획 수립, 대학 연구소의 학술연구에 활용되고 있으며, 일반 이용자에게는 기초 정보와 지식을 제공하는 역할을 합니다.

한국미디어패널조사

조사 목적
빠르게 변화하는 미디어 환경에서 우리나라 가구와 가구 내 개인의 미디어 소비가 중장기적으로 어떻게 변화하는지를 파악

조사 방법
가구 방문 일대일 면접조사

조사 대상
전국 5,109 가구 및 해당 가구의 만 6세 이상 가구원 약 12,000 여명을 2011년부터 추적조사

방송산업 실태조사

조사 목적
방송사업자의 종사자와 시설, 프로그램 제작과 수출입, 방송 편성 현황 등 방송산업 전반에 대한 실태 파악

조사 방법
인터넷 설문조사

조사 대상
과학기술정보통신부 ·방송통신 위원회에 등록된 방송사업자 전수조사

KISDI STAT 사이트 및 미디어 통계수첩 소개

미디어통계포털(KISDI STAT)

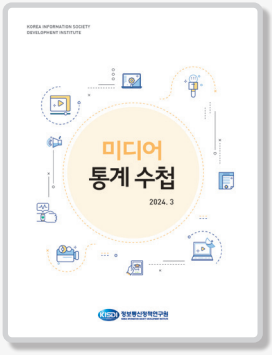
미디어통계포털(KISDI STAT)은 방송시장과 미디어 이용에 관한 다양한 조사 결과 데이터와 분석보고서를 편리하게 조회하고 활용할 수 있도록 만든 통계정보시스템입니다. 본 사이트에서는 한국미디어패널조사, 방송산업 실태조사, 해외방송통계 자료의 조건검색과 원시자료 다운로드가 가능하고, 최신 이슈를 데이터에 기초하여 분석한 KISDI STAT REPORT 등 다양한 분석자료가 제공됩니다.

KISDISTAT 미디어통계포털
http://stat.kisdi.re.kr

▲ 클릭하면 사이트로 이동

미디어 통계수첩(2024년 발간)

미디어 통계 수첩은 미디어 통계 이용자가 편리하게 활용하도록 돕기 위해 정보통신정책연구원의 한국미디어패널조사, 방송통신위원회의 방송매체이용행태조사, 과학기술정보통신부와 방송통신위원회의 방송산업실태조사의 주요 시계열 통계를 요약, 정리하여 작성한 소책자입니다. KISDI STAT 사이트에서 내려받을 수 있습니다.



▲ 클릭하면 사이트로 이동