**텍스트마이닝 분석을 통한 네트워크 분석과 경제 변수와 텍스트 데이터 관계 규명**

201801984 송찬영

1. **서론**

케냐는 동아프리카에 위치한 국가로 동쪽에는 인도양, 서쪽에는 빅토리아 호수와 인접해 있는 국가로, 석유와 같은 풍부한 천연 자원은 없지만, 농업과 관광업을 중심으로 1963년 독립 이후 1980년대까지 꾸준한 경제 성장을 이루었다.[[1]](#footnote-1) 대외경제정책연구원(KIEP)에 의하면 케냐는 1965년부터 1980년까지 경제 성장률 최고치 12.84% 평균 경제 성장률 5.25%를 기록하였지만, 1980년대 초반의 세계적 경기 침체, 1990년대 구조조정 프로그램 등의 원인으로 침체를 겪고 있다.[[2]](#footnote-2) 케냐공공정책연구원(Kippra)에 따르면 1달러 미만으로 생활하는 절대 빈곤층이 1997년 1,340만명에서 2006년에는 전체인구의 41%인 1,660만명으로 늘었다. 또한, 도시생활자 3명 중 1명이 월 40달러로 연명하고, 농촌지역 주민의 절반은 21달러로 생활하고 있다고 밝혔다.[[3]](#footnote-3) 이수철과 황재훈(2011)에 의하면 케냐 빈곤 문제의 원인은 경제 정책의 실패, 국가 역량 부족, 해외지원의 부족, 자본주의 시스템 구축 실패 등 다양한 원인이 지적되고 있다.[[4]](#footnote-4)

케냐의 경우처럼 국가 내에 자생적으로 해결할 수 없는 경제 문제의 경우, 국내에서 자생적으로 빈곤 문제를 해결하기 어려운 상황이다. 따라서, 대중 매체에서 케냐의 경제 문제가 어떻게 다뤄지고 있는지가 매우 중요하다. Scheufele, D. A.(2000)에 따르면 대중매체의 프레이밍과 의제가 어떻게 구성되는지에 따라 특정 이슈가 대중에게 인식되는 형태와 정도가 정해지고,[[5]](#footnote-5) 박재진과 김태우(2013)에 의하면 대부분의 사회 구성원은 사회문제와 같은 공익적 문제에 대해 그 문제점과 해결책을 충분히 인식함에도 만성적 불감증 때문에 행동으로 옮기지 못한다.[[6]](#footnote-6) 따라서, 케냐관련 뉴스 기사에서 케냐 경제 문제를 대외적으로 어떻게 다뤘는지에 따라 대중에게 케냐 빈곤 문제가 인식되는 형태와 심각 정도가 정해지고, 사회구성원이 해당 문제를 해결하기 위한 행동의 실마리를 제공하기 때문에 뉴스 기사에서 빈곤 문제를 어떻게 묘사하는지가 중요하다.

따라서, 본 연구에서는 2018년 케냐 관련 기사를 중심으로 어떤 키워드를 중심으로 네트워크의 형태를 파악하고, 뉴스 텍스트의 감정 분석을 통해 케냐의 경제 관련 변수가 어떻게 변화하는지 파악하는 것을 목표로 한다.

1. **데이터셋**

본 연구에서는 케냐 언론사인 Business Daily, Daily Nation, Standard 세 언론사의 날짜, 제목, 본문을 2018년 1월부터 2018년 12월까지의 기사를 대상으로 하여 크롤링한 뉴스 기사 데이터를 사용한다. 이후 회귀분석 시에는 PRS Country 데이터의 동기간의 국가 경제 관련 데이터를 사용한다. PRS 데이터는 총 43개의 변수 68,062 행으로 구성되어 있는 데이터로 1984년부터 월별 변수가 기록되어 왔다. PRS Country데이터에는 각 국가의 ‘Financial Risk Rating’, ‘Economic Risk Rating’, ‘Risk for Budget Balance’와 같은 경제 관련 변수가 있고, ‘Corruption’, ‘Civil Disorder’, ‘Democratic Accountability (K)’와 같은 정치, 사회 관련 변수도 내장되어 있다. 또한, 뉴스 기사 데이터를 통해 각 월별 감정(sentiment)과 anger, anticipation등과 같은 개별 감정 평균을 계산한 뒤, 이를 PRS의 월별 데이터에 적용하여 사용한다. 해당 연구에서는 2018년 1월부터 12월까지 12개 행을 대상으로 사용하였다.

1. **데이터 전처리**

기사 텍스트 데이터는 tidytext 패키지를 활용하여 토큰화 데이터 클린징을 하였다. 이 과정에서 숫자, 기호, 하이픈, 구두점, URL을 제거하였다. 이중 알파벳만 남겼고, 모든 문자를 소문자로 변환한 뒤, 어근을 추출하였다. 불용어는 기본적으로 제공된 불용어 사전을 활용하였고, develop, business와 같은 단어는 빈도가 높게 나타났지만, 해당 단어로 기사의 유의한 결과를 도출할 수 없다는 연구자의 판단으로 제외하였다.

1. **EDA**

**4.1 뉴스 기사 데이터**

**4.1.1 토큰 별 비중 그래프**

**텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명**

그림 토큰 별 비중

그림1은 2018년 1월 1일부터 12월 31일까지 1년 동안의 기사 본문의 주요 단어 등장율을 그래프화 한 것이다 .polit(politicians의 어근), public, trade, invest, support가 빈도율 상위 5개를 차지하고 있다.

**4.1.2 공기어 네트워크 분석**

2018년 1월부터 12월까지 전 구간의 공기어 네트워크의 형태를 살펴 본 후, 상반기(1~6월), 하반기(7~12월)로 나누어 살펴본다.

텍스트, 도표, 라인, 지도이(가) 표시된 사진

자동 생성된 설명

그림 2018년 공기어 네트워크

그림2는 2018년 전 구간의 기사의 공기어 네트워크 분석의 시각화 결과이다. Public과 invest를 중심으로 네트워크가 형성된 것을 확인할 수 있다. Public의 네트워크 형태로 현재 케냐의 공공 이슈 중에 부채(dept), 재정(finance), 직업(job), 부패(corrupt) 등이 연관되어 있는 것을 확인할 수 있다. 또한, elect와 polit(politicians의 어근)을 통해 2017년에 있었던 대통령 선거와 그 여파가 현재 진행 중인 것을 확인할 수 있다. 실제, 2013년 케냐타 정부에서부터 제기되어 왔던 부패문제는 심각한 상황이고, 해당 이슈가 2017년 대상 당시에도 제기되었고, 그 여파로 2018년에도 public과 polit(politicians의 어근)을 잇는 네트워크로 등장하였다.[[7]](#footnote-7)

텍스트, 도표, 라인, 지도이(가) 표시된 사진

자동 생성된 설명

그림 상반기 공기어 네트워크

상반기 공기어 네트워크에서는 polit(politicians의 어근)과 elect의 단어 중심으로 네트워크가 형성된 것을 확인할 수 있다. elect와 polit(politicians의 어근)에 corrupt가 밀접하게 형성되었다. 부패 관련 이슈가 상반기에는 주요하게 자리 잡은 것을 확인할 수 있다. 실제, 18년 상반기 당시 많은 정치인이 부패를 없애겠다는 정치인의 공약이 기사 내용을 확인할 수 있었다.[[8]](#footnote-8)

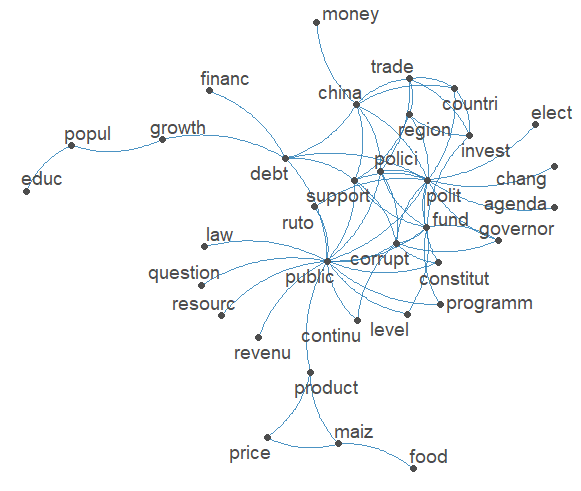


그림 하반기 공기어 네트워크

하반기에는 elect가 네트워크에 사라진 것을 볼 수 있다. 2017년 주요 이슈였던 대통령 선거가 하반기에서는 주요하게 나타나지 않았다. 대신, public, polici과 같은 공공, 정책 같은 이슈가 네트워크 내에서 주요하게 자리 잡았다.

**4.1.3 감정분석**

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

그림 기간별 감정 그림 감정별 빈도

그림 5는 기간별 감정을 시각화한 것이다. 긍정, 감정 모두 비슷한 기간에 증가폭이 상승하고 하락하는 것을 확인할 수 있다.

그림 6은 각 감정의 비율을 나타냈다. 긍정 감정에서 가장 높은 비율을 차지한 감정은 trust가 17.9%로 가장 높았고, anticipation이 10.0%로 뒤를 이었다. 부정의 감정에서 가장 높은 비율을 차지한 감정은 fear가 7.3%로 높았고, anger가 5.5%로 기록되었다.

**4.2 PRS 데이터**

**4.2.1 기술통계**

표 변수 별 기술통계량

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | mean | min | Q1 | Q2 | Q3 | max | std |
| Composite Risk Rating | 63.8125 | 63.25 | 63.25 | 64.0 | 64.0625 | 64.25 | 0.4281 |
| Corruption (F) | 1.75 | 1.5 | 1.5 | 1.75 | 2.0 | 2.0 | 0.2611 |
| **Economic Risk Rating** | **32.25** | **31.5** | **31.5** | **32.5** | **32.5** | **33.0** | **0.5839** |
| **Financial Risk Rating** | **37.75** | **37.5** | **37.5** | **37.75** | **38.0** | **38.0** | **0.2611** |
| **Investment Profile (C)** | **8.375** | **8.0** | **8.375** | **8.5** | **8.5** | **8.5** | **0.2261** |
| Political Risk Rating | 57.625 | 57.5 | 57.5 | 57.5 | 57.625 | 58.0 | 0.2261 |
| Repatriation | 2.875 | 2.5 | 2.875 | 3.0 | 3.0 | 3.0 | 0.2261 |
| **Risk for Budget Balance** | **4.0833** | **4.0** | **4.0** | **4.0** | **4.0** | **4.5** | **0.1946** |
| **Risk for Current Account as % of XGS** | **9.3333** | **9.0** | **9.0** | **9.5** | **9.5** | **9.5** | **0.2462** |
| **Risk for Debt Service** | **9.25** | **9.0** | **9.0** | **9.25** | **9.5** | **9.5** | **0.2611** |
| **Risk for Foreign Debt** | **6.1667** | **5.5** | **6.0** | **6.25** | **6.5** | **6.5** | **0.3892** |
| **Risk for Inflation** | **8.3333** | **8.0** | **8.0** | **8.5** | **8.5** | **8.5** | **0.2462** |

PRS데이터의 각 변수 별 기술통계량은 다음과 같다. 해당 표에 없는 Bureaucracy Quality (L), Civil Disorder 등과 같이 2018년 전구간 같은 값을 가지는 변수는 본 연구에서 제외되었다. 본 연구에서 사용할 경제 관련 변수로 선택한 것은 ‘**Economic Risk Rating’, ‘Financial Risk Rating’,** ‘**Investment Profile (C)’, ‘Risk for Budget Balance’, ‘Risk for Current Account as % of XGS’, ‘Risk for Debt Service’, ‘Risk for Foreign Debt’, ‘Risk for Inflation’** 이다.

**4.2.2 상관분석**

사각형, 스크린샷, 직사각형, 패턴이(가) 표시된 사진

자동 생성된 설명

그림 변수 별 상관계수 히트맵

그림 7은 PRS데이터의 각 변수 별 상관계수를 히트맵으로 시각화한 것이다. ‘Composite Risk Rating’과 ‘Economic Risk Rating’의 상관계수 0.93, ‘Composite Risk Rating’과 ‘Risk for Current Account as % of XGS’의 상관계수는 0.97, ‘Risk for Debt Service’와 ‘Financial Risk Rating’의 경우 상관계수가 -1까지 나타났다. 높은 상관계수가 나타나게 된 배경에는 해당 변수들이 밀접하게 연관이 있다는 점도 있지만, 2018년 단일 연도의 데이터만 사용하여 데이터의 수가 부족했고, 월별 데이터의 변동성이 적었다는 점이 높은 상관계수의 원인이 되었다.

1. **회귀분석**

**5.1.1 산포도 및 회귀선**

텍스트, 도표, 라인, 평행이(가) 표시된 사진

자동 생성된 설명

그림 sentiment x 경제관련변수 회귀선

그림 8은 sentiment와 주요변수로 설정한 8개의 변수와 산포도 및 회귀선이다. 'Economic Risk Rating', 'Risk for Budget Balance', 'Risk for Current Account as % of XGS', 'Risk for Inflation' 이 기울기가 양인 회귀선을 보여주고 있고, 'Risk for Foreign Debt'만 유일하게 음의 회귀선을 보여주고 있다. 외의 변수는 수평에 가까운 회귀선을 보여주고 있다.

**5.1.2 단일선형회귀 분석**

단일선형회귀 분석에서는 뉴스기사를 통해 산출한 감정변수인 ‘Sentiment’를 독립변수로 하고, 본 연구에서 정한 PRS의 각 경제 변수 7가지를 종속변수로 설정한다. 각 변수의 종류는 표2와 같다.

설정한 가설은 다음과 같다.

H0: Sentiment는 각 변수에 유의한 영향을 미치지 않는다.

H1: Sentiment는 각 변수에 유의한 영향을 미친다.

표 단일선형회귀분석 결과

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dependent Variable** | **Sentiment**  **Coefficient** | **Coefficient p-value** | **R-squared** | **Normality p-value** | **Independence p-value** | **Homoscedasticity p-value** | **ADF p-value** |
| Economic Risk Rating | 0.2472 | 0.2491 | 0.1303 | 0.0477 | 0.7557 | 0.7557 | 0.0047 |
| Financial Risk Rating | 0 | 1 | 0 | 0.0174 | 1 | 1 | 0 |
| Investment Profile (C) | 0.1201 | 0.8339 | 0.0046 | 0.0556 | 0.2098 | 0.2098 | 0 |
| Risk for Budget Balance | 0.7415 | 0.2491 | 0.1303 | 0.0392 | 0.349 | 0.349 | 0.0035 |
| Risk for Current Account as % of XGS | 0.4635 | 0.3686 | 0.0814 | 0.0278 | 0.9607 | 0.9607 | 0.2219 |
| Risk for Debt Service | 0 | 1 | 0 | 0.0174 | 1 | 1 | 0 |
| Risk for Foreign Debt | -0.1854 | 0.5746 | 0.0326 | 0.0588 | 0.7891 | 0.7891 | 0 |
| Risk for Inflation | 0.4635 | 0.3686 | 0.0814 | 0.0278 | 0.9607 | 0.9607 | 0.2219 |

모든 경우에서 Coefficient는 유의수준 0.05에서 귀무가설을 기각할 수 없다. R-squared 값은 ‘Economic Risk Rating’과 ‘Investment Profile (C)’에서 0.1303으로 가장 높게 나타났지만 전체적으로 낮은 R-squared 값을 가졌다. 정규성 가정은 ‘Investment Profile (C)’, ‘Risk for Foreign Debt’을 제외하고 유의수준 0.05 하에서 만족하였다. 독립성 가정과 등분산성 가정 모두 유의수준 0.05하에서 만족하였고, 시계열 데이터의 정상성은 ‘Risk for Current Account as % of XGS’, ‘Risk for Inflation’ 변수에서만 유의수준 0.05하에서 만족하지 않았다.

**5.1.3 다중선형회귀 분석**

‘Economic Risk Rating’을 종속변수로, 독립변수를 'Financial Risk Rating', 'Investment Profile (C)', 'Risk for Budget Balance', 'Risk for Current Account as % of XGS', 'Risk for Debt Service', 'Risk for Foreign Debt', 'Risk for Inflation', 'anger', 'anticipation', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'trust', 'negative', 'positive', 'sentiment'으로 본 경제 관련 변수와 감정 변수로 설정하였다. 설정된 가설은 다음과 같다.

H0: 독립변수들은 종속변수에 영향을 미치지 않는다. 회귀계수(β)는 모두 0이다.

H1: 적어도 하나의 독립변수는 종속변수에 영향을 미친다. 적어도 하나의 회귀계수(β)는 0이 아니다.

**텍스트, 번호, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

그림 다중회귀분석 결과

분석 결과 sentiment와 investment Profile (C)를 제외한 모든 변수가 유의수준 0.05에서 기각될 수 있다. F통계량 또한 유의수준 0.05에서 해당 회귀식이 유의하다고 볼 수 있다. Prob(Omnibus)의 경우 유의수준 0.05로 정규성 가정에 큰 문제가 없다고 할 수 있다. 하지만, Durbin-Watson 통계량은 0.172로 매우 작아 자기 상관이 있을 가능성이 높고, Cond. No가 매우 작으므로 다중공선성에 문제가 있다고 판단할 수 있다.

1. **결론 및 한계점**

본 연구에서는 2018년 동안의 케냐에 관한 뉴스 기사를 중심으로 네트워크 분석과 감정 분석을 수행하였으며, 뉴스 기사에서 얻은 감정 변수와 PRS 데이터의 경제 관련 변수들 간의 관계를 다중회귀분석을 통해 살펴보았다.

상반기와 하반기의 뉴스 기사에서는 부패, 대통령 선거와 같은 정치적인 이슈가 강조되었으며, 상반기에서는 부패와 정치 이슈가, 하반기에서는 공공 정책과 정부 정책에 대한 이슈가 중요하게 다뤄졌다. 기부의 활성화 방안으로 기부처의 투명성 확보가 중요한 요인인만큼[[9]](#footnote-9), 부패이슈가 자선단체나 개인의 기부 의도에 저해 요소로 작용하지 않기 위해선 부패 이슈의 해결로 언론 내의 부패 이슈 언급율을 줄일 필요성이 있다.

단일선형회귀 분석에서는 Sentiment가 경제 변수들에 유의한 영향을 미치지 않는 것으로 나타났다.

다중선형회귀 분석에서는 다양한 경제 변수들 중에서 Sentiment, Investment Profile (C)를 제외한 모든 변수가 유의한 영향을 미친다는 결과가 도출되었다. 그러나 자기 상관이 있을 가능성이 높은 Durbin-Watson 통계량과 다중공선성 문제가 있음을 나타내는 Cond. No 값이 확인되어, 해당 결과를 유의하다고 볼 통계적 근거는 없다.

본 연구에서 회귀분석 시 사용된 감정 변수는 경제 변수를 설명하는 데에 다중공선성의 문제가 있음을 확인하였다. 또한, 타 경제 관련 변수 역시 다중공선성의 문제가 존재하였다. 뉴스 기사에서의 감정 분석은 감성사전에 의존하여, 정확도와 일반화에 대한 한계가 있을 수 있다.

본 연구는 케냐의 경제와 관련된 뉴스의 감정과 경제 변수 간의 통계적 관계에 대한 초석으로서 의의가 있다. 2018년 케냐의 언론에서 주요하게 다뤄진 단어와 고이겅 네트워크 분석을 통해 그 관계를 확인하였다. 후속 연구에서는 2018년 단일 연도의 데이터가 아닌 10년 이상의 데이터를 활용하여 데이터의 변동성을 더 파악할 수 있는 토대를 마련하여 진행한다. 또한, 변수간 다중공선성의 문제를 해결한 연구를 기대한다.

1. Logovini, A. (2000). *Kenya: Macro Economic Evolution Since Independent*. World Bank [↑](#footnote-ref-1)
2. KIEP. (2023). [전문가오피니언] 케냐의 청년실업 위기: 원인 분석과 정책 해법 제언 [↑](#footnote-ref-2)
3. 우만권, “케냐 국민, 10년 전보다 더 가난해져”, 한겨레, 2009.10.13. [↑](#footnote-ref-3)
4. 이수철, 황재훈. (2011). 케냐 빈곤의 정치경제학적 조망. 지역발전연구, 20(2), 25-51. [↑](#footnote-ref-4)
5. Scheufele, D. A. (2000). Agenda-setting, Priming, and Framing Revisited: Another Look at Cognitive Effects of Political Communication. Mass Communication & Society, 3(2-3), 297-316. [↑](#footnote-ref-5)
6. 박재진·김태우(2013), 공익광고의 메시지 프레이밍 효과-해석수준이론을 중심으로, 언론학연구, 17(2), 63-87. [↑](#footnote-ref-6)
7. KIEP(2018), 케냐타 2기 정부의 출범 배경 및 향후 정책 과정 [↑](#footnote-ref-7)
8. James Giahuye(2018). Facing a "broke" country, Liberia's Weah cuts his own salary, Reuter. [↑](#footnote-ref-8)
9. 서희열, 심충진, 조영탁(2008), 기부금 활성화를 위한 투명성 확보방안, 세무학연구 제25권 제2호. [↑](#footnote-ref-9)