

텍스트 기반 모델링의 경험

조남경¹

1. Nam Kyoung Jo

한국디지털사회복지학회 편집
분과위원장, 성공회대학교 시
민사회복지대학원장
namk.jo@skhu.ac.kr

국문 초록과

영문 제목 및 영문 초록은 생략합니다.

I. 들어가며

근거와 논리로 주장하고 설득하는 논문에 당연히 분석자의 생각과 ‘마음’이 다 담길 수 없다. 본격 학술지의 모습은 아직 아닌 ‘디지털과 사회복지’는—물론 나중에 ‘한국학술지인용색인(KCI)’에 등록하더라도, 복지와 기술의 만남에 대한 현장과 학계의 경험과 고민을 나누는 것을 목표로 하기에 보통의 학술지와는 다른 모습이 기대되지만—솔직하게 경험을 털어 놓고 나눌 수 있는 기회를 준다고, 내 마음대로 생각하기로 했다.

2022년부터 2023년까지 2년 동안 사회복지와 연관성이 있는 세 가지 서로 다른 내용의 비정형 텍스트 자료를 분석하는 기회가 있었다. 사회복지 연구에서 아직도 현장의 실제 자료를 빅데이터 분석해 볼 기회는 많지 않다. 텍스트 자료라면 더더욱 그렇다. 그만큼 운이 좋았다는 뜻이다. 분석자로서 나는 무엇을 하였고 어떤 생각들을 했는가.

II. 무엇을 했나

첫 번째 분석자료는 통합돌봄 선도사업을 해 온 광주 서구가 보유하고 있는, 집에 거주하고 계시는 75세 이상 어르신을 방문하고 나서 방문자가 메모한 기록으로, 2년 반이라는 기간에 걸친 익명 처리된 20만여 건의 규모있는 자료였다. 여기에 방문기록이 있는 어르신들 중 일부는 통합돌봄의 대상자로 선정되어 서비스를 제공받았고, 이 또한 결합시킬 수 있는 자료로 존재하고 있었다. 그렇기 때문에 인공지능의 지도학습 방법—결과값에 해당하는 정보가 있어서 어떤 특성들(을 가진 사례들)이 어떤 결과값과 상관성이 높은지를 학습하게 하는 방법—을 활용하여, 방문 기록이라는 텍스트 자료(가 가진 특성)만을 가

지고 어떤 어르신이 통합돌봄을 제공받은 분이고 어떤 어르신이 아닌지를 맞춰보는 모델을 만들어 볼 수 있겠다 생각했다. 상당한 정도의 확률로 맞힐 수 있다면, 앞으로는 방문기록만으로도 대상자 분류에 대한 예측이 가능하다는 것 아닌가. 물론 그 ‘상당한 정도의 확률’, 즉 판별/예측의 정확도가 문제다. 통합돌봄 대상자 선정과 관련하여 ‘선별도구’라 불리는 ‘필요도 조사’가 개발되어 있는데, 이 전통적인 설문 방식의 조사에 의한 선별보다 정확하지 않다면 큰 의미를 갖기는 어렵다. 결과는 나쁘지 않았다. 80%가 넘는 예측 정확도를 보였고, 이는 ‘필요도 조사’에 의한 모델의 예측 정확도보다 10% 포인트 이상 높은 것이었다.¹⁾

두 번째 분석자료는 2년 동안 서울시 50플러스재단의 4개 캠퍼스를 찾아 대면, 전화, 온라인의 방법으로 상담한, 익명 처리된 약 2만여 건의 상담 기록이다. 규모가 상대적으로 작아 머신러닝 방법으로 빅데이터를 분석하는 장점이 드러나기 쉽지 않았다고 생각되었지만, 어떤 결과값 정보를 가지고 지도학습을 시켜서 예측 모델을 만드는 것이 아니라, 비지도 학습의 토픽(topic) 모델링 방법을 활용하여 상담 기록으로부터 참여자들의 일자리와 관련된 관심과 욕구를 추출해보는 분석의 기회를 놓칠 수는 없었다.²⁾ 데이터 규모가 충분히 크지 않아 토픽들 간 차이점이 아주 선명하지는 않은 결과에 만족해야 했지만, 자격증 취득을 통한 취업의 욕구, 요양보호사나 사회복지사로의 취업 욕구, 디지털 기기 활용 역량 배양 욕구, 창업 관련 관심 등을 포함한 12개의 토픽을 추출할 수 있었다.³⁾

세 번째 분석자료는 익명 처리된 5년 간 약 24만 건의 경찰청 변사사건 현장감식 자료였다. 같은 자료를 가지고 전국 고독사 통계를 처음 생산해 본 선행연구가⁴⁾ 고독사

여부를 판별한 결과값 정보를 결합시킬 수 있었기 때문에, 앞서 통합돌봄 관련 분석 사례에서와 마찬가지로, 지도학습에 의해 고독사 여부를 판별하는 모델을 만들어 보는 것을 계획했다. 역시나 나쁘지 않은 예측 정확도의 모델을 도출할 수 있었는데, 문제는 이 분석의 목표는 고독사 통계 작성을 위해 고독사 여부 판정을 빠르게 할 수 있는 과정을 만드는데 기여하는 것이기 때문에 ‘상당히 좋은 예측력’은 큰 의미가 없다는 점이었다. 덕분에 규모 있는 비정형 텍스트 자료를 대상으로 열셋말(keywords) 분석을 하는 다양한 아이디어를 개발하고 이를 자료의 문맥 속에서 확인하고 해석하는 좋은 훈련이 되었다. 결과적으로는 극히 일부인 고독사 사례를 헤아리기 위해 하나하나 검토해야 하는 전체 변수 자료의 15~20% 정도를 줄여줄 수 있는, ‘고독사가 아닌 변수’와 연관된 열셋말들을 제시할 수 있었다.⁵⁾

III. 무엇을 느꼈나

‘잘 정제된 질 좋은 소규모 데이터보다 들쭉날쭉 엉망인 대규모 데이터가 훨씬 좋은 결과를 낸다’는, 빅데이터 전환을 가져오게 한 금언은 사실이었다.⁶⁾ 전통적인 (추론)통계방법에 익숙한 관점에서는 머리로는 이해되지만 믿기는 어려운 일이다. 방문기록 자료를 처음 받아보고는 적지않게 실망했었다. 있는 그대로의 어르신인 상태나 느낌, 생각 등이 담기기보다 방문의 증거를 남겼다는 성격이 강한 기록—예를 들면 무엇을 가져다 드렸고 어떤 당부를 드렸다는—들이 대부분인 것처럼 보였다.⁷⁾ 따라서 분석결과를 보고는 ‘이런 결과가 어떻게 가능하지?’ 싶었다. 내 스스로 과정 하나하나를 이해해보는 논문을 썼던 이유다.⁸⁾ 글에 등장하는 모든 낱말과 낱말들의 조합을 변수로 삼아 상관성을 분석하다니, 참 ‘징그럽게 무식한’ 방법인데, 그만큼 높은 예측력도 수궁이 간다. (그리고, 컴퓨터에서 최신의 게임을 하지 않는 사람이라면 소위 ‘286 컴퓨터’ 시절 이래로 컴퓨터 성능의 발전과 변화를 그리 실감한 적은 없을 것이다. 나도 그랬다. 그런데 이 ‘무식한 인공지능 방법의 분석’을 빨리도 해 내는 것을 보면 실감하게 된다.) 이후의 분석 사례들도 한 목소리로 텍스트 빅데이터 분석의 가능성을 이야기해 주었다. 뚜렷한 목적을 가지고 개발된 설문 조사보다 텍스트에 의한 머신러닝 모델이 돌봄 필요도를 더 정확히 예측할 수 있다. 특별히 계획된(구조화된) 질문에 의한 것이 아닌 상담의 기록들로부터 대상자들이 어떤 생각들을 갖고 있는지 주요 토픽⁹⁾을 추출할 수 있다. 어느 정도 규모의 텍스트 자료만 있다면 대부분의 경우, 대상자를 분류하거나 대상자들이 갖고 있는 주요 욕구나 생각을 분석해 낼 수 있다는 이야기다. 심지어 녹음된 내용을 자동으로 텍스트

로 바꿔주는 기술이 상용화되어 있으니, 텍스트 자료의 질은 더 좋아질 일만 남았고, 분석 결과의 질도 더 정교해질 일만 남았다.

사회복지 현장에서의 활용 가능성도 확실히 체감되었다. ‘어떻게 할 수 있는가’는 더 이상 문제가 아니다. ‘무엇에 활용할까’가 문제이다. 이용내역 빅데이터 분석으로 각종 추천 서비스를 원하는 원치 않은 경험하고 있고, ‘피 검사’ 없이 하루에 텔레비전은 몇 분 보는지, 출퇴근은 무엇으로 몇 분 걸려 하는지, 세 끼는 주로 무엇을 먹는지 등을 묻는 것만으로 보험료를 산정하고, 별도의 신용정보 조회 없이 SNS 활동 내용 등을 가지고 신용 등급을 매기고 대출을 해 준다는 이야기도 듣지만, 사회복지와는 시간과 거리가 (아직은) 있는 일이라 여겼다. 그런데 나 같은 일개 연구자, 그것도 공학자도 개발자도 아닌 ‘문과’ 연구자도 할 수 있다니. 그렇다면 정말 누구나 할 수 있다. 명색이 연구자인 나는 분석 방법의 원리와 과정을 설명할 수 있어야 하지만, 소위 ‘논문쓰고 가르치는 일’이 없다면 생성형 AI에게 필요한 코드를 짜달라고 하면 된다.¹⁰⁾ 내 기관/조직에 있는 어떤 텍스트 자료를 가지고 이용자 분류 예측 모델을 만들어 볼지 혹은 이용자의 주요 생각/욕구들을 추출해 볼지 마음대로 상상해보면 된다. 상상했다면 내 기관에 있는 그다지 특별할 것도 없는 컴퓨터에서 분석해 볼 수 있다. 한 번 해 보면 그 다음은 자료만 다를 뿐 같은 분석이니 더 쉽다. 그러니 우리가 던질 질문은 ‘무엇에 활용할까’이다. 이 질문에 대한 답은 당연히 현장(의 경험과 고민)에서 나온다. 우리 기관/조직에서는 어떤 텍스트 자료를 계속 축적하고, 그에 기반한 어떤 모델을 만들어서 정기적으로 기관/조직의 활동에 의미있는 참고 자료를 생성하도록 할 것인가.

혹시 데이터 손질에만 몇 주 이상 걸릴 것 같다는 생각이 스멀스멀 올라오고 있다면 앞서 보았던 이야기를 반복해 읽자: ‘잘 정제된 질 좋은 소규모 데이터보다 들쭉날쭉 엉망인 대규모 데이터가 훨씬 좋은 결과를 낸다.’ 이 말은 텍스트 자료의 규모가 된다면 데이터 손질 걱정은 할 필요가 없다는 의미도 된다. 솔직히 나 스스로가 예전 방식의 통계 분석에 익숙한 연구자로서 데이터 정제를 매우 중요시하기에 맞춤법이 틀렸거나 띄어쓰기가 잘못된 경우를 그냥 보고 넘기기가 어려웠다. 이 글에서 언급한 세 가지 분석 사례 모두에서 그런 정제 작업을 시도하지 않은 적이 없었다. 그저 시간 낭비였다. (앞으로는 그러지 않을 것이다.) 예를 들어 면담기록을 하나하나 보면서 데이터 정제 작업을 한다면, 하루에 몇 백 건 하기도 어렵다. 데이터가 숫자가 아닌 비정형의 텍스트, 그냥 자연어로 된 문장들이기 때문에, 규칙을 만들어 수정하고 싶어도 명백한 한계가 있다. (그리고 그런 ‘수작업’ 방식

으로 데이터 정제를 할 수 있는 규모라면 아마도 빅데이터가 아닐 것이다!) 더구나 빅데이터 분석을 위한 자연어 처리 도구들은 이런 문제들의 상당히 많은 부분을 자동으로 해결해 준다. 더 중요한 것은, 빅데이터 분석에서는 내 눈에 보이는 자료의 불완전함은 극히 일부일 뿐이며 전체 분석 결과에 아주 미미한 영향만을 미칠 뿐이라는 것을 이해만 하는 것이 아니라 믿는 것이다.

IV. 나오며

결론을 써야 하는 글이 못되니 그야말로 ‘글을 나오며’ 사족을 보탠다. 이해는 하지만 믿지는 못하고 있는 문제를 두 번이나 언급했는데, 인공지능의 방법을 활용하여 빅데이터 분석을 수행하는 사람으로서의 솔직한 어려움이다. 빅데이터 ‘시대’라 불리는 이유는 그 전 시대와 질적으로 다른 점들이 있기 때문이며, 따라서 그 시대에 적합한 관점 혹은 사고 방식으로의 전환이 필요하다.¹¹⁾ 그래야 ‘빅데이터적인’ 질문, ‘인공지능의 방법 활용에 적합한’ 질문을 더 많이 던지고 그 답을 찾는 새로운 분석들을 해 볼 수 있을 텐데, 그게 참 어렵다. 그런데, 직접 해 보는 경험은 크게 도움이 되었다.

(신념의 의미가 아니라 보다 완전한 이해의 의미로 사용한 것이지만) ‘믿음’을 언급하는 바람에 혹시나 ‘기술 신봉자’로 오해받을까—물론 그 반대로 오해받는 것도—두렵다. 우리가 다소 부정적인 의미로 ‘기술 신봉자’와 같은 말을 만들어 사용하는 이유는 대개 ‘기술은 사람 같지(인간적이지) 않다’는 생각 때문이다. 이런 생각이 기술(기계, 인공지능, 로봇)이 인간을 지배하거나 인간과 대립하는 디스토피아를 상상하는 바탕인데, 철학자 김진석은¹²⁾ 그것이 얼마나 논리적으로 취약한 ‘인간중심주의’적인 생각인지 파헤쳐 보여준다. (그의 책은 철학적 깊이가 있는 것이지만 딱 내 수준으로 이야기하면, 가장 비인간적인 혹은 ‘짐승 같은’ 짓을 저지르는 것도 인간인데 ‘인간적’이라는 기준이 무엇인가.) 하지만 그가 일깨워 주는 더 중요한 점은 다른 곳에 있다. 미래의 대립은 인간의 일자리를 뺏는 기술과 인간 사이가 아니라 인간 대신에 기술을 투입하는 인간과 인간 사이에서 일어난다는 점이다. 인간과 기술의 대립을 부각시키는 것은 기술을 소유하고 활용하여 이익과 권력을 축적할 수 있는 사람들과 그렇지 못한 사람들 사이의 권력관계나 지배관계가 여전히 핵심일 수밖에 없음을 잊게 만든다. 단지 일자리의 문제만이 아니다. 우리가 ‘사람처럼 자율적으로 판단하고 행하는’ 기술을 추구할수록—이것도 사람처럼 하는 것이 최고라는, 우리의 인간중심주의 때문이 아닐까—네트워크된 체계가 필요하고 인간이 자율성을 행사할 여지는 삭제될 수 있다. ‘자율’주행차는 도로, 목적지, 가능한 경로들,

보행자, 다른 차량 등 주행과 관련한 모든 조건과 상황 등이 실시간으로 자율주행차와 정보를 주고받을 수 있는, 자율주행을 위한 만물인터넷(IoE, Internet of Everything) 체계 같은 것이 갖춰질 때 가능하며, 그 때 인간은 기계-알고리즘-체계로 연결된 ‘고리의 바깥’에 놓이게 된다. 자율성은 시스템을 관리하고 통제하는 권력에만 허락된다. (이 때 개인이 차를 소유하고 관리하는 일은 시스템에 의한 관리와 통제를 벗어나는 ‘위험한’ 일이 될 것이라는 지적은 소름 끼친다.) 알고 있다고 생각하던 지적이 새삼 크게 다가오는 이유다: “인공지능과 관련된 모든 정책과 개발에 시민이 발언권을 가져야 한다. (중략) 새로운 기술은 사회를 좀 더 민주적이고 평등하게 할 수도 있고, 반대로 이미 많은 권력을 가진 사람들에게 더 많은 권력을 가져다줄 수도 있다.”¹³⁾

너무 거창하게 되어 버렸지만 하려던 이야기는 사실 (한국디지털사회복지 학회의 회원들에게는) 뻔한 것이다. 기술이 사회복지와 대립적인 것이 아니라, 기술을 적용하는 방식과 목적이 사회복지와 대립적일 수 있다. 우리는 더 나은 사회복지를 위해 기술과의 접목을 고민하지만, 기술이 인간과 인간 사이의 불평등을 강화하고 약자와 소수자의 소외를 심화시키지 않도록 ‘깨어 있기’ 위해서도, 사회복지의 영역에서 기술 도입과 관련하여 올바른 방향으로 시민의 발언권을 지키고 행사하기 위해서도 기술을 이해하고 실제적 적용 경험을 쌓을 필요가 있다. 물론 기술 자체는 가치중립적이지 않아서 근대 이후의 기술은 근대 합리주의 및 원자론적 개인주의와 닿아 있고, 스스로의 독립적 자율성을 향해 중단없이 나아간다. 아무런 통제와 제약이 없다면 중국에는 인간의 자율성이 배제될 수 있는 이유다. 하지만, 적절히 통제되며 활용된다면 인간 다수의 자율성을 증진하고 강화하는 수단이 될 수 있다. 언제 어떻게 어떤 방식과 방향으로 통제되고 활용되어야 하는가? 해 본 사람이 알 수 있다. 막상 해 보니 별것도 아닌 텍스트 빅데이터 인공지능 모델링, 당장 해 봐야 하는 이유다.

Notes

- 1) 조남경, 조재성, 남일성, 송기호, 손다진, 2022, 광주 서구 통합돌봄 모니터링 및 성과분석 연구: 대상자 빅데이터 분석을 중심으로, 광주광역시 서구·성공회대학교 산학협력단.
- 2) 이 분석에서는 대표적이고 가장 많이 쓰이고 있는 토픽 모델링 방법인 잠재 디리클레 할당(LDA, Latent Dirichlet Allocation) 방법을 사용하였다.
- 3) 이석환, 조남경, 장익현, 2023, 서울시 보람일자리 사업 모델 개발 연구, 서울시50플러스재단 연구보고서 2023-006.

- 4) 고숙자, 안영, 황남희, 이아영, 최현수, 2023, 2022년 고독사 예방 실태조사 연구, 보건복지부·한국보건사회연구원 정책보고서 2023-32.
- 5) 가칭 ‘고독사 판단기준 개선방안 연구’로, 언급하고 있는 분석은 전체 연구의 한 부분이며 조남경과 조재성이 함께 작업하였다. 연구보고서는 아직 나오지 않았다.
- 6) 마이어쉴버거, 쿠키어 저, 이지연 역, 2013, 빅데이터가 만드는 세상: 데이터는 알고 있다, [Mayer-Schönberger, V. and Cukier, K., 2013, Big Data: a revolution that will transform how we live, work, and think], 경기 파주: 21세기북스. / 물론 이러한 전환은 대규모 데이터의 분석을 가능하게 해 준 컴퓨팅 기술의 발전이 전제되었기에 가능한 것이었다.
- 7) 이것은 선입견이었다고 부분적으로 인정해야 할 것 같다. 빅데이터 시대에 텍스트 자료를 분석하면서 일반적인 통계 분석에서 몸에 익은 습관인 ‘데이터 전체에 대한 감을 잡고 시작’하는 것은 선부른 일이기 쉬웠다. 화면에서 글자 크기를 내 눈이 허락하는 가장 작은 수준으로 줄여 놓고 방문기록을 보면 한 번에 한 30개 정도 된다. 인쇄심을 가지고 읽어 나가도, 내 경우에는 1,000개 정도 읽는 것이 최대치였다. 그 다음에는 마우스를 붙잡고 잡아 내리거나 올리면서 눈을 사로잡는 특정한 패턴이 보이면 살펴보게 되는데, 대부분은 짧고 건조한 한 문장이 기록된 사례들이 모여 있어서 갑자기 화면에 여백이 많이 등장하거나 그림같은 패턴을 만드는 경우였다. 그러니 처음 수백 개의 자료에서 받은 부정적 느낌은 ‘스캐닝’하듯 훑어보는 과정에서 계속 강화되었다. 몇 번인가 비슷한 과정을 되풀이했지만, 전체 20여 만개에서 1만개도 못 보는 이 과정은 선입견을 갖게 하는 것 외에 별로 좋은 점이 없었던 듯하다. 내용을 구체적으로 살피는 일 없이 분석해야 한다는 주장이 아니다. 어차피 분석 과정에서 열쇳말들의 문맥적 의미를 알기 위해 그 낱말이 포함된 사례들을 추출해서 살펴보는 일을 반복한다. 기존의 통계 분석 때처럼 데이터 전체에 대한 어떤 느낌과 이해를 갖고 시작하려 하는 것이 규모있는 텍스트 자료를 다루는 과정에는 잘 안 맞는 것 같다는 이야기를 하는 것이다.
- 8) 조남경, 송기호, 2023, 사회복지의 상담기록, 좀 더 활용할 수 있을까? ‘머신러닝’을 통한 사회복지 상담 텍스트 활용 가능성의 점검, 한국사회복지조사연구 79: 5-26.
- 9) 굳이 ‘토픽’이라고 말하는 이유는, 토픽 모델링이라고 하는 방법을 통해 추출되는 토픽들을 바로 ‘주제(theme)’라고 이해하기는 어려운 측면도 있기 때문이다.
- 10) 지금 이 순간에 네이버 클로버X에게 “상담기록에서 명사만 추출해서 잠재 디리클레 할당 방법으로 토픽모델링을 하는 파이썬 코드를 짜줘”라고 요청해서 순식간에 답을 받았다. 코드가 영어 자연어 처리 기준으로 작성되었다는 점이나 추출할 토픽을 5개로 정해놓았다는 점 정도 말고는 크게 아쉬울 것 없는 명료한 코드였다. 몇몇 대중적인 파이썬 코드 작성기 프로그램들은 이미 생성형 AI를 옵션으로 붙여 사용할 수 있게 하고 있다.
- 11) 조남경, 2019, 질적, 양적 연구를 넘어? 사회복지 빅데이터 연구방법의 모색, 한국사회복지학 71(1): 7-25.
- 12) 김진석, 2019, 강한 인공지능과 인간, 경기 파주: 글항

아리.

- 13) 장정일, 2024, 인공지능과 민주주의, 녹색평론 185(2024년 봄호): 32-41 중 40-41쪽.