데이터로 인해 발생하는 자연어처리 분야의 윤리적 이슈

강혜린°, 장연지◇, 강예지, 박서윤, 김한샘†
연세대학교 언어정보연구원†, 국립국어원◇

{hyerink°, yjkang5009, seoyoon.park, khss†}@yonsei.ac.kr, yeonji3547@korea.kr◇

Ethical Issues in Natural Language Processing arising from Data

Hyerin Kang°, Yeonji Jang◇, Yejee Kang, Seoyoon Park, Hansaem Kim†
Institute of Language and Information Studies, Yonsei University†
National Institute of Korean Language◇

요 약

자연어처리에서 데이터는 굉장히 많은 부분을 차지하고 중요한 역할이지만, 데이터로 인한 윤리적 이슈 또한 많이 나타난다. 본 연구는 자연어처리에서의 데이터 흐름의 과정에서 나타날 수 있는 윤리적 이슈를 단계별로 정리하였다. 이는 복잡한 자연어처리 과정의 특성과 자연어처리 분야에서 나타나는 상황을 모두 고려한 것이다. 또한 단계별로 정리한 이슈를 토대로 자연어처리가 더 나은 방향으로 나아가기 위한 데이터 관점에서의 미래 방향을 제시하였다.

주제어: 자연어처리에서의 윤리, 인공지능 윤리, 데이터 윤리

1. 서론

그동안 자연어처리 분야는 비약적으로 발전했으나 자연어처리 과정에서 발생할 수 있는 윤리적 문제에 관한 국내적 논의는 부족하였다. 윤리는 옳고 그름을 분별할 수 있도록 해준다. 자연어처리의 기술적 성장에 집중하는 동안 그 과정에서 발생할 수 있는 여러 옳고 그름에 관한 기준이나 문제에 대해서는 다소 관심을 기울이지 않았다. 여러 모델을 개발하고 이를 활용한 기술적 성취는 우리 생활 전반에 긍정적 영향을 끼쳤으나 윤리적인 고려 없이 행해진 결과 사회에서 수용될 수 없는 결과를 보여주기도 하였다. 이는 자연어처리에 있어서 윤리적 기준에 부합하는 개발이 필수적임을 보여주는 결과라 할 수 있다. 자연어처리는 기술 집약적이며 과정이 매우 복잡하여 결과를 내기까지 상당히 많은 단계를 거쳐야 한다. 그만큼 윤리적 문제의 종류도 다양하고 복잡한 양상을 보인다. 따라서 자연어처리의 각 과정을 따라가면서 과정마다 발생할 수 있는 이슈를 유형화하여야 한다. 자연어처리는 데이터로 무엇을 하는지에 따라 여러 단계로 나눌 수 있다.

자연어처리에 있어 데이터는 매우 필수적이며 중요한 역할을 수행한다. 데이터 없는 자연어처리는 불가능하다. 그만큼 데이터는 자연어처리에 있어 중요하나 그런 만큼 자연어처리의 전체 과정에 있어 데이터로 인한 문제도 발생한다. 이러한 문제 중 일부는 윤리적 이슈를 가진다.

본 연구는 데이터 활용 흐름의 관점에서 자연어처리의 각 과정에서 발생할 수 있는 윤리적 이슈를 살펴본다. 또한 윤리적 이슈를 해결하기 위한 앞으로의 방향성을 모색하고자 한다.

2. 관련 연구

[1]에서는 자연어처리에서 윤리적 문제를 다루는 연구를 크게 세 개 분야로 구분한다. 첫 번째는 자연어처리 분야에 전통적 윤리학을 적용하는 연구이다. 아리스토텔레스, 밀, 칸트 등 윤리학에 큰 영향을 미친 학자의 논의를 자연어처리 윤리에 적용하는 방식이다. 두 번째는 빅데이터, 데이터마이닝, 머신러닝 등에서 제기되는 윤리적 함의에 관한 컴퓨터 과학 분야의 연구이다. 이는 편향성, 차별, 개인정보와 같은 문제를 중심으로 연구가 진행된다. 마지막으로는 컴퓨터 교육과 관련하여 직업윤리의 문제를 다룬 연구이다. 본 연구에서는 두 번째 분야를 중심으로 자연어처리 과정에서 발생하는 윤리적 이슈에 관한 문헌을 중심으로 검토하고자 한다.

2.1 자연어처리에서의 윤리적 이슈

자연어처리에서 데이터로 인해 나타날 수 있는 중요한 윤리적 이슈로 다뤄지는 것 중 하나는 편향이다. 임베딩(embeddings), 언어 모델링(language modeling), 대화 생성(dialogue generation) 등 다양한 자연어처리 태스크에서 편향의 결과가 나타난다. [2]에서는 GPT-3 와 같이 단어 임베딩을 학습하는 거대 언어 모델은 더 높은

수준의 교육을 나타내는 능력과 직업을 남성에 우선적으로 연관시킨다고 하였다. 임베딩에서 나타나는 성별 편향성은 자연어처리의 학습 데이터에서 나타나는 직업과 비율이 모델에서 영향을 받는다[3]. 이러한 편향은 잠재적으로 사회적 편견을 강화할 위험이 있다. [4]는 자연어처리의 과정에서 나타나는 5가지의 편향의 원인에 관해 설명한다. 자연어처리 파이프라인에서 나타나는 편향은 데이터 주석을 시작으로 데이터 선택, 데이터 표현, 모델과 연구 설계(research design)에 이르기까지 모든 과정에서 나타난다. 따라서 이러한 편향을 해결하기 위한 방법을 제시하였는데, 데이터의 균형 유지와 편향 제거를 위한 기술적 접근이 필요하다 말하였다. [5]는 편향을 주제로 한 100 여 개의 논문을 분석한 결과 편향성을 가지는 시스템의 결과가 어떤 방식으로 누구에게 부정적인 영향을 미치는지에 대한 규명이 부족하다고 지적하였다. 이는 편향에 대한 개념화가 명확하지 않아 나타난 현상이며 편향을 분석하기 위한 새로운 접근법을 제공한다.

자연어처리 윤리에서 또 하나의 중요한 이슈는 프라이버시이다. 프라이버시가 포함된 데이터는 여러 윤리적 이슈를 가져온다. [6]은 데이터 수집을 위하여 API 를 통해 twitter 데이터를 크롤링할 때에 함께 수집되는 Personally Identificable Information(PII)가 사용자에게 더 나은 서비스를 제공하는 데에 도움이 되지만, PII 가 악용된다면 사용자에게 심각한 피해를 줄 가능성이 있다고 하였다. [7]은 머신러닝 기반의 인공지능 시대에 프라이버시가 큰 관심사이며 머신러닝 모델과 훈련 데이터셋이 모두 개인정보 보호에 있어 위험한 요소이며 민감한 정보가 유출될 가능성이 있다고 하였다. [8]은 자연어처리 분야 다수의 연구가 웹에서 크롤링한 자료를 적극적으로 활용하지만 민감한 데이터가 어떻게 처리되었는지를 자세히 서술한 논문이 거의 없다는 것을 지적하였다. 데이터를 활용한 연구를 정량적으로 분석하여 데이터의 수집, 저장 및 배포에 대한 내용을 살펴본 결과 연구 과정에서 민감한 데이터를 수집하거나 사용한 연구 중 3.5%의 연구만이 데이터의 익명화에 관해 서술하였다고 한다.

2.2 자연어처리에서의 윤리적 이슈 해결을 위한 기준

민감한 데이터 사용으로 인해 나타날 수 있는 윤리적 이슈를 최소화하기 위해서는 데이터 사용에 있어 책임 의식이 필요하다. [9]에서는 자연어처리에서 책임감 있는 데이터 사용을 위해 기준으로 삼을 수 있는 체크리스트를 제공하였다. 자연어처리 윤리 관련 사항의 핵심 부분은 연구자의 책임감 있는 데이터 사용이지만 이에 관한 정확한 정의와 방향성에 대한 내용이 명확하지 않음을 지적하면서 데이터 사용에 있어 참고할 수 있는 체크리스트를 마련하였다. 체크리스트는 이미 공개된 리소스를 사용하는 경우와 새로운 리소스를 사용하는 경우에 따라 나누었다. 이미 공개된 리소스를 사용하는 경우에는 데이터를 선택하는 데에 있어 리소스의 대표성 문제 등의 제한 사항과 데이터 보호 문제를 모두 고려했는지를 확인하여야 한다. 또한 license 에 대한 부분도 필수적으로 확인되어야 한다고 정리하였다. 새로운 리소스를 사용하는 경우는 리소스를 만드는 과정부터 시작이 된다. 논문에서 데이터 수집에 관련한 전체 과정을 자세히 서술하여야 하며 사람의 노동력이 투입된 주석 작업의 경우, 공정한 보상을 지급하여야 한다고 명시한다. 또한 데이터를 안전하게 사용을 강조하였다. 데이터의 안전한 사용을 위해서는 데이터의 출처와 기본적인 데이터 비율에 관한 설명을 상세히 서술하여야 하며 민감 정보 데이터에 있어 피해를 보는 사람이 발생하지 않도록 익명화에 노력을 기울여야 한다고 말한다.

[10]에서는 개발자와 사용자를 위한 인공지능용 윤리 가이드라인을 마련하였다. 규범적 내용을 명확화하여 유형별로 권장 사항을 명시하였다. 이는 인공지능에 관련한 분야에서 나타날 수 있는 모든 이슈를 포괄하고 있다. 윤리적 원칙(ethical principle)을 대분류로 구성하고 윤리적 원칙별로 세부 윤리적 이슈 요소로 가이드라인의 큰 틀을 마련하였다.

3. 자연어처리 과정에서의 윤리적 이슈

본 연구에서는 자연어처리의 전반적인 과정을 데이터의 흐름과 연결 지어 접근하고, 단계별 윤리적 이슈를 유형화하여 살펴본다.

자연어처리에서 데이터는 데이터 수집, 데이터 가공, 데이터 컨트롤, 데이터 표현의 과정을 거친다. 데이터 수집은 자연어처리 태스크에 적합한 데이터 종류를 찾고 해당 데이터를 가져오는 과정을 말한다. 데이터 가공의 과정에서는 데이터를 태스크에 적합하게 전처리나 목적에 맞게 주석 작업을 수행한다. 데이터 활용 과정에서는 가공한 데이터를 모델에 적용하여 활용한다. 구체적인 기술적 방법론이 적용되는 단계이다. 데이터 표현은 자연어처리의 최종 결과물로 모델의 아웃풋을 말한다.

3.1 데이터 수집

자연어처리 태스크에 맞는 데이터를 수집하는 과정은 자연어처리 전 과정에 있어 출발점이자 중요한 단계이다. 태스크의 목적과 자연어처리로 얻고자 하는 결과와 직접적으로 연관된 데이터를 수집해야 한다. 이미 공개된 기존 리소스를 사용할 수도 있으나 연구의 목적에 맞는 데이터셋을 직접 구축하기 위해 연구 시작과 동시에 데이터를 수집하는 경우도 많다. 이러한 데이터 수집 과정에서는 고르지 못한 데이터를 수집함으로써 나타나는 편향성 문제와 프라이버시 관련 문제가 나타난다.

3.1.1 데이터 수집 과정에서의 편향 발생

태스크 목적에 맞는 데이터를 수집하는 과정에서 데이터 전체의 균형성이 맞지 않게 일부만을 수집한다면 데이터는 편향성을 가지게 된다. 수집한 데이터는 일부 집단의 언어 사용 양상만을 보여주는 경우가 존재한다. 이는 어휘 사용의 양상이나 데이터 내에서 드러나는 의미적 관계가 한정적일 수 있음을 나타낸다. 또한 수집한 데이터는 한정된 도메인만을 대상으로 할 수도 있다. 이러한 경우 데이터는 다양한 경우를 반영하지 못한다. 편향의 결과 데이터 상에서 배제되는 부분이 생겨나며 이는 연구의 보편성과 객관성을 위협하는 요소로 작용할 수 있다. 데이터 편향은 이후 모델을 학습시키고 학습 결과를 확인하였을 때도 그대로 반영되기 때문에 중요한 윤리적 이슈로 다루어져야 한다.

3.1.2 데이터 수집 과정에서의 프라이버시

자연어처리의 태스크가 다양한 만큼 수집되는 데이터의 유형과 도메인도 다양하다. 수집되는 데이터의 유형이 많은 만큼 데이터를 이루고 있는 정보는 다양하다. 사실이나 상식 정보를 포함하는 경우도 있으나 개인의 정보를 표현하는 데이터가 존재할 수 있다. 수집하는 데이터에 개인에게 민감하게 작용할 수 있는 프라이버시 정보가 포함될 경우 이로 인한 윤리적 이슈가 나타날 수 있다. 데이터 도메인의 특성을 파악하지 않고 데이터를 수집한다면 프라이버시 문제에 직면할 수 있다. 또한 데이터 유형마다 포함하는 개인 정보의 유형이 다르게 나타나는데, 유형의 특성에 맞는 프라이버시 문제를 고려하지 않는다면 다량의 개인정보는 가져오고 극히 일부의 개인정보를 수집하지 않는 상황이 발생할 수 있다. 연구 목적에 맞는 데이터를 수집하기로 결정했다면 해당 도메인에서 나타나는 다양한 개인정보 포함 여부 및 개인정보의 출현 형태를 확인하여야 할 것이다.

3.2 데이터 가공

데이터 수집 과정이 완료되었다면 이후 연구 목적에 맞게 데이터를 가공하여야 한다. 가공은 데이터를 전처리 및 연구 목적에 맞는 주석 작업 수행으로 이뤄진다. 주석 작업은 데이터의 정교함을 위해 작업자의 노동력이 개입되는 경우가 빈번하다.

3.2.1 데이터 가공 과정에서의 프라이버시 처리

데이터 전처리 과정에서는 개인의 정보가 포함된 데이터의 익명화 작업이 중요하다. 개인 데이터의 오용이나 도덕적 및 법적 영향을 미칠 수 있기 때문이다. 자연어처리의 응용 분야 중 특히 상호작용이 필수적인 태스크에서는 비윤리적 표현의 유무를 탐지하고 비윤리적인 표현을 분류하는 것이 매우 중요한 부분이다. 따라서 혐오, 공격성 등을 담은 문장을 수집하여 연구 데이터셋을 구축하는 경우가 있다. 주로 댓글이나 트위터를 대상으로 데이터를 하는데, Twitter 의 tweet 을 연구 데이터로 수집한 경우[11][12] user id 는 개인 정보로 포함해 비식별 처리가 된 데에 반해 tweet 내에 나타난 개인 정보에 대해서는 비식별 처리나 익명화가 이뤄지지 않았다. 논문 내용에도 문장 내 개인 정보에 해당하는 부분을 어떻게 처리하였는지에 관한 서술이 나타나지 않는다. 비윤리적 표현의 대상에 관한 보호가 중요하지만, 데이터를 가공하는 과정에서 이러한 요소를 고려하지 못한 것으로 보인다. 개인을 특정할 수 있는 정보나 개인에게 피해가 갈 수 있는 정보가 데이터에 포함된 경우에 데이터 가공 과정에서 익명화가 제대로 이루어지지 않는다면 이후 피해가 발생할 가능성이 있다.

3.2.2 투명성

데이터 가공 과정에서의 투명성은 앞서 언급한 프라이버시와 연결된다. 민감한 정보가 어떻게 처리되었는지에 관한 투명한 공개가 필요하나 이러한 부분이 부족한 상황이다. 데이터 가공 과정에 관한 투명한 공개가 보장되지 않는다면, 이는 연구 참여자의 주관이 개입될 가능성도 있다. 편향 발생의 위험이 있는 경우는 어떻게 처리하였는지에 대한 정보도 공개될 필요가 있다. 데이터 가공 과정에서 나타나는 여러 윤리적 이슈는 결국 투명성이 선행되지 않아 나타난 결과라 할 수 있다.

3.2.3 크라우드소싱

대량의 데이터를 연구 목적에 맞게 가공하는 데에는 많은 시간과 비용이 요구된다. 특히 데이터에 주석 작업을 진행할 경우에는 많은 사람의 노동력이 동원된다. 원 데이터에 주석 작업을 진행하거나 기구축된 데이터를 검수하여 수정하는 작업도 크라우드소싱 과정에 포함된다. [13]에서는 자연어처리 연구에서 크라우드 워커를 통한 데이터 생산이 빠르게 증가하고 있으나 크라우드 워커에 관한 윤리적 논의가 급여나 노동 조건으로 범위가 제한된다고 지적하였다. 노동력이 동원되는 작업이지만 크라우드 워커가 수행하는 작업에 관한 윤리적 고려가 부족한 상황이다. 또한 크라우드 워커를 통해 데이터를 가공한 연구의 경우 급여에 관해 언급한 논문은 20% 내외였으며 Institutional Review Board(IRB)의 승인을 받은 경우는 2020 년 크라우드소싱을 사용했다고 발표한 180 개의 논문 중 5 개에 불과하였다. 이는 크라우드소싱을 연구 과정에 활용할 때 윤리적 고려와 접근이 부족한 것으로 해석할 수 있다.

인력이 동원되는 만큼 그 과정에서 올 수 있는 다양한 변수를 고려한 윤리적 기반이 없다면 이는 크라우드소싱

과정에서 무엇이 옳은지 그른지에 관한 분별이 부족할 가능성이 있다.

3.3 데이터 활용

데이터 가공 과정을 통해 연구 목적에 맞는 데이터를 구축하였다면 이를 모델에 활용하여야 한다. 데이터 활용은 주석한 데이터로 모델을 훈련시키거나 기술적인 부분의 조정을 하는 단계이다. 이 과정에서 훈련 데이터의 편향성으로 인한 모델의 편향이 윤리적 이슈로 나타날 수 있으며 자연어처리의 데이터 의존적 특성으로 인한 이슈가 발생할 수 있다.

3.3.1 데이터 활용 과정에서의 편향

데이터 수집 과정에서 나타난 편향이 보완되지 않고 데이터 활용 과정에 이르게 된다면, 이는 자연스레 모델의 편향에도 영향을 미치게 된다. 편향된 데이터가 입력으로 들어온다면 모델도 편향적인 모델이 될 것이다. 모델은 훈련 데이터에 존재하는 편향을 그대로 받아들일 것이기 때문에 사회적 편견이나 유해한 편향이 결과에 그대로 반영될 가능성이 높다. 이러한 기술적 이슈가 발생할 위험성이 존재한다.

3.3.2 데이터 의존성

바로 위에서 언급한 모델의 편향은 자연어처리의 데이터 의존성으로 인해 생겨난 결과이다. 자연어처리에 활용되는 모델은 데이터가 필수적이다. 즉, 모델은 모델 자체만으로는 어떠한 문제를 해결할 수 있는 능력이 없다.

자연어처리가 발전할수록 사람은 자연어처리 응용 분야를 통해 인공지능과 상호작용하는 일이 점점 늘어나고 있다. 이런 상호작용의 결과로 만들어진 텍스트 중 일부는 다시 모델을 만드는 데에 사용된다. 유해한 데이터가 재사용되는 과정을 거친다면 더욱 유해한 모델이 만들어질 가능성이 있는 것이다. 이러한 모델의 데이터 의존적인 특성을 파악하고 데이터 수집을 시작으로 데이터 활용에 이르는 과정까지 데이터를 다루는 데에 유의해야 할 것이다.

3.4 데이터 표현

데이터 표현은 앞선 세 단계를 거쳐 모델이 최종적으로 만들어 낸 결과를 확인하는 단계이다. 즉, 이는 아웃풋에 해당하는데 자연어처리 최종 결과물에서 나타날 수 있는 윤리적 이슈는 편향, 자연어처리 특성에서 오는 복잡성, 결과물에 관한 책임 소재에 관한 문제가 있다.

3.4.1 데이터 표현 과정에서의 편향

이전 과정에서 수정되지 못한 편향은 모델의 아웃풋에 그대로 담기게 된다. 이전 과정에서의 편향 필터링이 제대로 이루어지지 않은 경우 아웃풋은 직접적으로 결과가 드러나 그 편향의 문제가 확연히 느껴지게 된다. 또한 자연어처리 응용 분야를 사용하는 사용자에게 이러한 결과물이 노출된다면 부정적인 영향을 미칠 수 있다. 편향으로 인한 차별적 메시지 등을 전달할 가능성도 존재한다. 문제의 소지가 있는 결과물이 공개된다면 사회 전반에 있어 악영향을 줄 위험이 있다.

3.4.2 복잡성

자연어처리는 상당히 복잡한 과정을 거치기 때문에 결과로 나온 것에 대해 왜 이런 결과가 나타났는지를 논리적으로 설명할 수 없다. 자연어처리의 복잡다단함 때문에 데이터 결과에 관해 설명이 불가하다. 자연어처리로 인해 생겨나는 결과에 대해 설명이 필요한 때이다. 자연어처리 과정과 모델의 복잡성은 결과물로 인해 생겨날 수 있는 문제에 대한 책임 회피의 가능성과도 연결된다.

3.4.3 책임성

위에서 언급한 대로 자연어처리의 복잡성은 책임성과 연관된다. 자동화된 모델이 만들어낸 결과이기에 직접적으로 책임이 없다는 식의 책임 회피를 할 수 있게 만든다.

자연어처리결과물에 관한 책임성을 다소 약하게 만드는 것은 IRB 의 면제이다. 현재까지도 자연어처리 관련 연구는 IRB의 승인 절차가 면제된다. 자연어처리가 IRB의 승인 절차 면제 대상이 된 데에는 연구 대상인 데이터가 이미 공개되어 있기 때문에 이로 인해 발생할 수 있는 문제가 미미하다고 간주한 것에 있다. 하지만 인간의 상호작용, 특성에 관련한 모든 것이 데이터에 담겨 있는 지금은 수집한 데이터로 이전에는 분석이 불가능했던 것을 분석할 수 있다. SNS 에서 생산된 데이터를 연구에 직접 활용하게 되면서 연구 결과의 영향력이 커졌다고 볼 수 있다. 또한 이러한 상황은 사람이 실험 결과로 인해 직접적으로 영향을 받을 수 있다는 것을 의미하기도 한다[14]. 따라서 자연어처리의 책임 소재를 명확하게 하기 위한 제도적 마련이 필요한 상황이다.

4 자연어처리에서의 윤리적 이슈 해결을 위해 나아가야 할 방향

지금까지 살펴본 윤리적 이슈로 인한 문제가 생겨나지 않기 위해서 자연어처리가 어떠한 방향으로 나아가야 하는지를 각 데이터 흐름의 단계별로 제안점을 제시한다.

4.1 데이터 수집

자연어처리 태스크 목적에 맞는 데이터를 수집하는 과정에서 나타날 수 있는 윤리적 이슈는 편향과 프라이버시 관련 사항으로 정리할 수 있다.

이러한 윤리적 이슈로 인한 문제 상황을 방지하고 데이터 수집 과정이 올바르게 이뤄지기 위해서는 데이터의 대표성과 균형성이 중요하다. 데이터가 언어 현상을 최대한 대표하여 담고 있는지를 고려해 데이터를 수집해야 할 것이다. 또한, 데이터가 언어 사용 양상의 전반적인 모습을 골고루 담고 있는지도 고려의 대상이다. 연구의 목적에 따라 모든 언어 사용 양상을 데이터 내에 담을 수는 없겠지만 균형성을 염두에 둔다면 편향적인 데이터 수집은 조금이나마 방지할 수 있을 것이다. [14]에서는 데이터를 연령과 지역 편향적으로 수집할 경우, 수집된 데이터가 응용 분야에 적용되었을 때 일부 사용자가 소외될 것이라는 점을 지적하면서 데이터 수집 단계에서부터 지역별, 연령별 화자 분포를 고려해야 한다는 점을 강조하였다. 이러한 균형성을 기억하고 데이터를 수집하여 데이터셋을 구축하여야 할 것이다. 이는 연구를 설계하는 과정에서 수행되어야 한다.

또한, 데이터 수집 과정에서 데이터 내에 포함될 수 있는 개인정보에 대한 사전적 파악이 필요하다. 도메인 내에서 표현될 수 있는 개인정보의 유형에 대한 사전 학습이 필요하다. 더욱 다양한 종류의 개인정보가 나타남에 따라 데이터를 수집할 때 민감한 정보로 취급하여 하는 대상이 늘어났다. 다양한 도메인 유형별 개인정보의 출현 사항에 관한 정리가 필요한 때이다.

4.2 데이터 가공

데이터 가공 과정에서는 가공 과정에서의 부족한 익명화 작업으로 인해 발생할 수 있는 프라이버시 침해 문제와 이와 연관되는 투명성 그리고 크라우드소싱 관련 윤리적 이슈가 발생할 수 있음을 살펴보았다. 데이터 수집 과정에서 마련된 개인정보 관련 가이드라인을 통해 해당 데이터 내에 포함된 모든 개인정보를 익명화 처리해야 할 것이다. 또한, 연구 내용의 서술 과정에서 민감한 정보의 처리 방식을 상세히 보여주어야 한다. 현재 국립국어원 '모두의 말뭉치'에 공개된 '메신저 말뭉치(버전 2.0)' 와 '온라인 대화 말뭉치(버전 1.0)'는 말뭉치 내 개인의 신원이 노출될 우려를 반영하여 비식별화 기본 지침을 마련하였으며 이에 따라 개인정보 비식별화 작업을 진행하였다. 또한 도메인의 특성을 반영하여 메신저 내의 기능도 개인정보 비식별화 대상이 되어야 함을 지침에 명시하였다. 또한 결과 보고서 내에 민감한 개인정보를 처리하고 비식별화하는 과정을 검수 단계별 작업 내용을 통해 상세히 서술하였다. 이러한 개인정보의 중요성을 기억하고 데이터 가공 시 민감하게 다뤄져야 하는 개인정보를 철저하게 처리하고 이 과정에 대해 투명하게 공개하여야 할 것이다.

더불어 크라우드소싱 관련 윤리적 기반을 마련하여 인력이 동원되는 과정에서 나타날 수 있는 여러 문제 상황을 모두 고려해야 할 것이다. 이는 합리적인 주석 작업 지불 비용을 포함하여 작업 전반에 관한 윤리적인 기준 마련이 필요해 보인다. 그리고 크라우드소싱 관련 내용도 연구와 밀접하게 연관되는 사항으로 연구 진행 사항에 자세히 서술하여야 할 것이다.

4.3 데이터 활용

가공까지 완료한 데이터는 데이터 활용 과정을 거치는데, 이 단계에서는 편향과 모델의 데이터 의존성에 관련한 윤리적 이슈가 발생한다. 모델의 편향을 완화하기 위해서는 데이터 수집 과정에서부터 균형 있는 데이터를 구축하여야 한다. 또한, 모델 수립 및 활용 과정에서 데이터 활용의 결과를 수시로 점검하여 편향을 조절해야 할 것이다. 여러 데이터 계층에 대한 모델을 훈련한다면 모델 편향을 어느 정도 제거할 수 있으리라 생각된다. 데이터 수집에서부터 엄격한 기준을 가지고 연구를 진행해 나간다면 자연어처리 모델의 데이터 완전 의존성에 올 수 있는 문제점을 어느 정도 해결할 수 있다. 모델의 뛰어난 성능도 중요하지만, 최종 단계에 이르기 전 혹은 연구 내용을 최종적으로 공개하기 전까지 모델의 중간 결과를 계속해서 확인하는 비판적인 자세를 가져야 할 때이다.

4.4 데이터 표현

자연어처리 최종 단계인 데이터 표현에서는 편향, 복잡성, 책임성 관련 윤리적 이슈의 출현 가능성이 있음을 살펴보았다. 이전 과정에서 편향이 제거되거나 완화되는 과정을 거쳤다면, 데이터 표현 단계에서의 편향에 관련한 윤리적 문제 상황이 발생할 가능성은 거의 없다고 볼 수 있다. 마지막 단계에 이르기까지 데이터를 세심하게 다루었다면 최종 단계에서 모델은 그에 상응하는 결과물을 보여줄 것이다. 이를 위해서는 모델의 최종 결과를 정량적으로 평가하는 데에서 더 나아가 정성적 평가를 통해 모델의 한계를 파악하고 이를 데이터에 보완하는 선순환적 조치가 필요하다.

또한 자연어처리의 모델의 결과에 대해 연구자는 책임 의식을 가지고 자연어처리 전 과정을 다루어야 할 것이다. 결과에 대한 설명이 최대한으로 가능하도록 하기 위해 자연어처리 모델의 데이터 활용 측면에서의 논리성을

확보해야 할 때이다. 이를 위해서는 제도적으로 자연어처리 분야에서 연구 성격에 따라 IRB 승인에 관한 기준 마련이 필요하다.

5. 결론

자연어처리가 더욱 발전하고 상용화됨에 따라 자연어처리에서의 윤리가 점점 더 중요해지고 있다. 자연어처리에서 데이터는 굉장히 중요하지만, 데이터로 인한 윤리적 이슈도 많이 나타난다. 본 연구는 자연어처리 내의 데이터의 흐름의 과정에서 나타날 수 있는 윤리적 이슈를 규명하고자 하였다. 데이터 수집, 가공, 활용, 표현에 이르기까지 각각의 과정에서 문제를 야기할 수 있는 윤리적 이슈를 찾아 정리하였다. 또한, 이러한 이슈를 발판 삼아 더 나은 자연어처리로 나아가기 위한 데이터 관점에서의 미래 방향을 제시하였다.

향후 연구에서는 지금까지의 연구에서 정리한 내용을 토대로 하여 윤리적 이슈 해결을 위한 가이드라인 마련 연구를 진행할 예정이다. 더불어 데이터를 활용하는 기술적인 부분에서 윤리적 이슈를 해결할 수 있는 방안을 고안하고 이에 대해서 향후 실험 연구를 통해 검증할 예정이다.

참고문헌

[1] Leidner, Jochen L., and Vassilis Plachouras. "Ethical by design: Ethics best practices for natural language processing." Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. 2017.

[2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D, “Language models are few-shot learners”. Advances in neural information processing systems, 33, 1877-1901, 2020.

[3] Toney-Wails, Autumn, and Aylin Caliskan. "Valnorm quantifies semantics to reveal consistent valence biases across languages and over centuries." arXiv preprint arXiv:2006.03950, 2020.

[4] Hovy, Dirk, and Shrimai Prabhumoye. "Five sources of bias in natural language processing." Language and Linguistics Compass 15.8: e12432., 2021.

[5] Blodgett, Su Lin, et al. "Language (technology) is power: A critical survey of" bias" in nlp." arXiv preprint arXiv:2005.14050 ,2020.

[6] Sohail, Shahab Saquib, et al. "Crawling Twitter data through API: A technical/legal perspective." arXiv preprint arXiv:2105.10724, 2021.

[7] Liu, Bo, et al. "When machine learning meets privacy: A survey and outlook." ACM Computing Surveys (CSUR) 54.2, 1-36, 2021.

[8] Mieskes, Margot. "A quantitative study of data in the NLP community." Proceedings of the first ACL workshop on ethics in natural language processing. 2017.

[9] Rogers, Anna, Tim Baldwin, and Kobi Leins. "Just What do You Think You're Doing, Dave?'A Checklist for Responsible Data Use in NLP." arXiv preprint arXiv:2109.06598, 2021.

[10] Ryan, Mark, and Bernd Carsten Stahl. "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications." Journal of Information, Communication and Ethics in Society ,2021.

[11] Kang, TaeYoung, et al. "Korean Online Hate Speech Dataset for Multilabel Classification: How Can Social Science Aid Developing Better Hate Speech Dataset?." arXiv preprint arXiv:2204.03262, 2022.

[12] Zampieri, Marcos, et al. "SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020)." arXiv preprint arXiv:2006.07235, 2020.

[13] Shmueli, Boaz, et al. "Beyond fair pay: Ethical implications of NLP crowdsourcing." arXiv preprint arXiv:2104.10097, 2021.

[14] 김진웅, “자연언어처리에서 윤리적 문제와 해결 방안: 연령 및 지역 편향성 극복의 출발점으로서 방언 자료 수집”. 연구방법논총, 6(1), 157-180, 2021.