기계 학습의 임상 텍스트 데이터에 대한 체계적 검토

Clinical Text Data in Machine Learning: Systematic Review

Irena Spasic, PhD and Goran Nenadic, PhD
JMIR Medical Informations. 2020 Mar

이동건
인하대학교 정보통신공학과
E-mail: time@inha.edu

2022. 05. 16.

- KEY WORDS:

  natural language processing

  machine learning

  medical informatics

  medical informatics applications

Contents

I. INTRODUCTION

기계 학습의 문제점 1

- 과거의 지식 도출 (knowledge elicitation) 병목 현상에 있어서, 기계 학습은 묘책 (silver bullet solution)으로 환영 받아 왔음

- 그러나, 기계 학습 모델의 훈련 데이터는 주석 처리 작업을 수반

- 많은 양의 데이터는 지식 도출 자체만큼 주석 작업에 많은 시간이 필요할 수 있음

기계 학습의 문제점 2

- 임상 서술의 유효성(the availability of clinical narratives)
  - 건강 데이터 및 개인 정보 보호 문제의 민감한 특성을 감안
- 수동으로 주석을 추가한 데이터를 사용할 수 없음
  - 학습 데이터의 대표성(representativeness)이 부족해짐
- 결과적으로 표준 이하 성능을 초래할 수 있음

주요 목표

- 임상 NLP에 대한 기계 학습 접근법을 훈련하는 데 사용되는 데이터의 특성에 대한 체계적인 증거를 제공

- 머신 러닝이 지원하는 NLP 작업의 유형과 임상 실습에 적용할 수 있는 방법을 조사

II. METHODS

Overview - 체계적 검토를 위한 단계별 방법론

- 1. 연구 질문(research questions, RQs) 사용
  - 검토의 범위, 깊이, 전체적인 목표를 규정하기 위해
- 2. 검색 전략 설계 (designing a search strategy)
  - RQ 관련 모든 연구를 효율적이고 재현 가능한 방식으로 식별하기 위해
- 3. 포함 및 제외 기준 정의
  - 범위를 세분화하기 위해
- 4. 포함된 연구에 대한 비판적 평가 수행
  - 검토 결과가 유효한지 확인하기 위해

Research Questions

Table 1 Research questions.

ID	RQ[a]
RQ1	What are the key properties of data used to train and evaluate machine learning models?
RQ2	What types of NLP[b] tasks have been supported by machine learning?
RQ3	How can NLP based on machine learning be applied in clinical practice?

RQ1	기계 학습 모델을 훈련하고 평가하는 데 사용되는 데이터의 주요 속성*은 무엇인가?
RQ2	기계 학습은 어떤 유형의 NLP 작업을 지원하는가?
RQ3	기계 학습 기반 NLP가 임상에 어떻게 적용될 수 있는가?

* 속성 : 크기, 출처, 이질성(내용, 구조, 임상영역)

검색 전략 (Search Strategy)

- PubMed를 검색 엔진으로 사용
- 리뷰의 주제를 설명하기 위한 검색어 목록 도출
  - *1)machine learning, 2)deep learning, 3)text, 4)natural language, 5)clinical, 6)health, 7)health care,* and *8)patient*
- 검색 전략에 따라 389개의 후보 글을 식별

((“machine learning”[All Fields] OR “deep learning”[All Fields]) AND (text[Title/Abstract] OR “natural language” [Title/Abstract]) AND (clinical[Title/Abstract] OR health [Title/Abstract] OR healthcare[Title/Abstract] OR patient [Title/Abstract]) NOT (literature[Title/Abstract] OR bibliometric [Title/Abstract] OR “systematic review”[Title/Abstract]) AND (“2015/01/01”[PDat] : “2018/08/08”[PDat])

```
(
        (
                "machine learning"[All Fields]
                OR
                "deep learning"[All Fields]
        )
        AND
        (
                text[Title/Abstract]
                OR
                "natural language" [Title/Abstract]
        )
        AND
        (
                clinical[Title/Abstract]
                OR
                health [Title/Abstract]
                OR
                healthcare[Title/Abstract]
                OR
                patient [Title/Abstract]
        )
        NOT
        (
                literature[Title/Abstract]
                OR
                bibliometric [Title/Abstract]
                OR
                "systematic review"[Title/Abstract]
        )
        AND
        (
                "2015/01/01"[PDat] : "2018/08/08"[PDat]
        )
)
```

검색 전략 (Search Strategy)

- 3~5번 줄: *1)machine learning*과 *2)deep learning*은 이 방법론들을 사용하는 기사를 검색하는 데 사용
- 9~11번 줄 : *3)text, 4)natural language* 는 학습 방법에 대한 관련 입력 유형을 나타냄
- 15~21번 줄 : 마지막 4개 용어는 임상 적용을 참조하는 데 사용
- 9~21번 줄 : 마지막 6개 용어의 광범위한 특성과 공통된 사용으로 인해, 제목(Title)과 초록(Abstract)으로만 언급 제한
- 23번 줄 : 비원본 연구 및 NLP 응용 프로그램의 검색을 방지하기 위해 *문헌 (literature), 서지학 (bibliometric)* 및 *체계적 검토 (systematic review)* 라는 용어들은 부정 처리
- 33번 줄 : 기계 학습의 새로운 응용에 초점을 맞추기 위해 검색기간을 2015년 1월 1일부터로 제한. 검색은 2018년 8월 8일에 수행됨

선택 기준 (Selection Criteria)

- Textbox 1. 포함 기준 (Inclusion Criteria)

1. NLP를 사용한 연구이어야 한다.
2. 이러한 처리를 지원하려면 기계 학습을 사용해야 한다.
3. 입력 텍스트는 의료 서비스 범위 내에서 일상적으로 수집되어야 한다.
4. 입력 텍스트는 써지거나(written) 받아써진 것(dictated)이어야 한다.
5. 동료 검토(peer review)를 받은 글이어야 한다.
6. 전문(full text)은 온라인에서 자유롭게 학술적으로 사용할 수 있어야 한다.

- Textbox 2. 제외 기준 (Exclusion Criteria)

1. 영어 이외의 언어로 작성된 글
2. 영어 이외의 언어의 자연어 처리
3. 구어의 자연어 처리

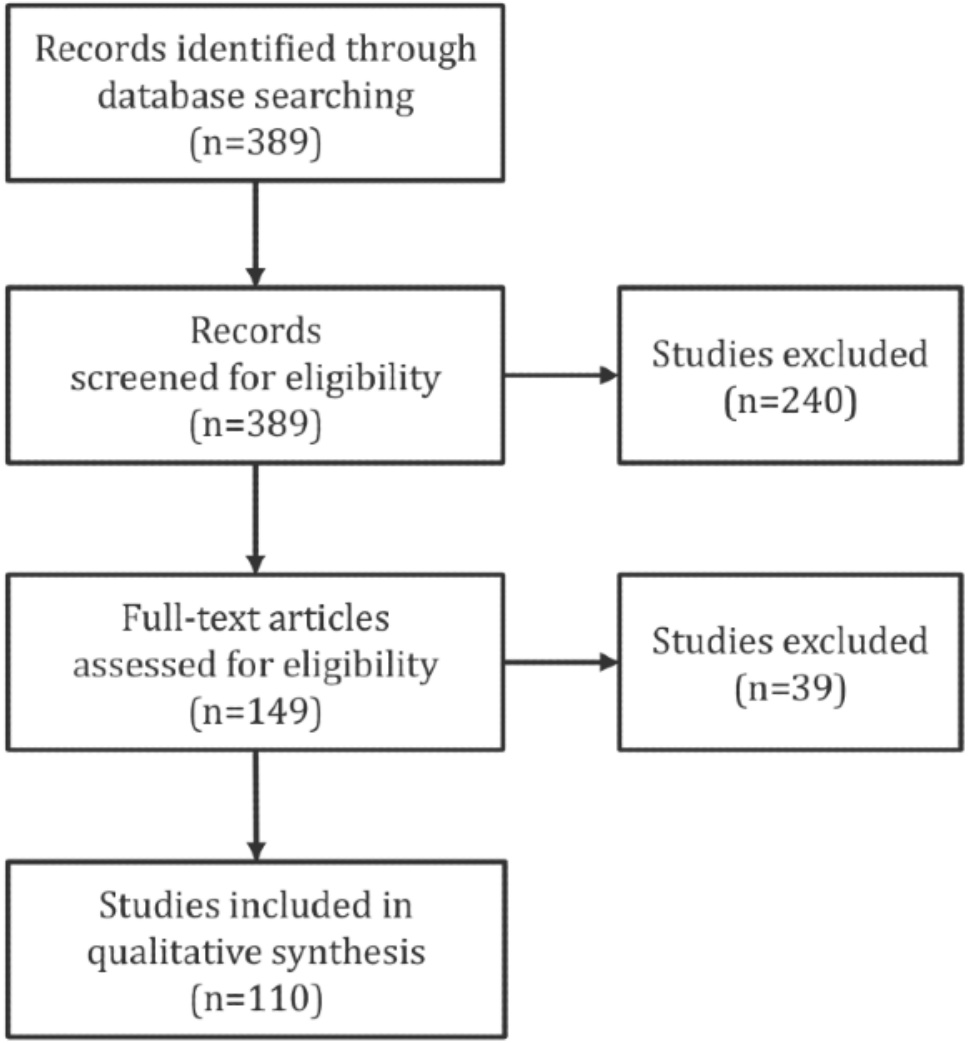
선택 기준 (Selection Criteria)

- 포함 및 제외 기준 (Textbox 1. 2.)
  - 검색된 389개의 글을 포함 및 제외 기준에 따라 149개로 선별

- 추가로, RQ에 응답하기 위한 충분한 정보를 제공해야 한다.
  - 사용된 데이터 세트를 설명하고,
  - NLP 문제를 명확하게 정의하며,
  - NLP를 지원하는 데 사용되는 기능을 설명하고,
  - 사용된 기계 학습 방법과 적절한 경우 매개 변수를 설명하고,
  - 결과에 대한 공식적인 평가를 제공해야 한다.

선택 기준 (Selection Criteria)

Figure 1. Flow diagram of the literature review process.



포함 및 제외 기준에 의거

추가기준에 의거

III. RESULTS

- Size
- Annotation
- Provenance
- Types of Narratives
- Clinical Applications

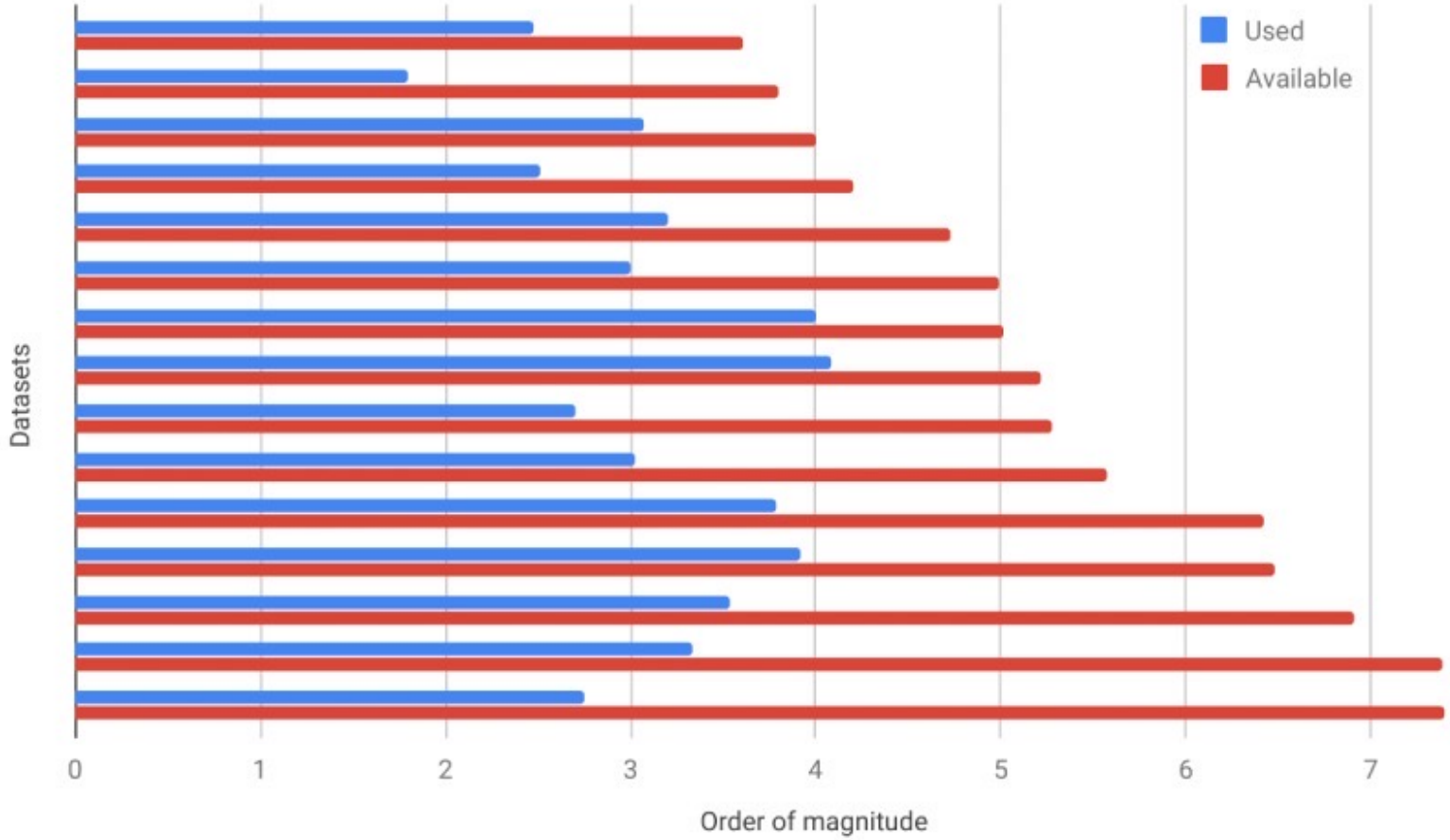
Size

- 훨씬 더 큰 데이터 세트를 사용할 수 있는 경우에도
  상대적으로 작은 데이터 세트가 활용되었음

- 사용 가능한 데이터의 0.002%만 활용하는 일부 연구 확인 가능
  (최대는 11.88% 활용)

Size

Figure 3. Data utilization on a logarithmic scale.



Annotation - 애로사항

- 데이터 활용률이 저조한 주된 이유?

- 기계 학습의 지도학습 (supervised learning) 알고리즘이 겪는 주석 병목현상 (annotation bottleneck) 때문
  - 지도학습 알고리즘은 데이터를 예측 수학 모델로 일반화하려면 훈련 데이터에 주석을 달아야 함

- 인간이 일일이 주석처리 하는 것은 노동 집약적(labor-intensive)이고 오류가 발생하기 쉬움

Annotation - 해결방안

- 1. 활성 학습 (Active Learning) 알고리즘 적용 [20, 54, 100]
  - 예측 모델의 품질에 따라 성능이 달라짐
  - 모델의 재학습이 상대적으로 오래 지속되는 경우 비효율적

- 2. 다양성 측정(diversity measure) 기법을 사용하여 주석의 우선 순위 지정
  - e.g. 코사인 유사도 기법
  - 이상값(outlier)이 존재하면, 모델의 성능 저하를 초래할 수 있음

참고 – Active Learning의 배경

- 1. 모델 학습 시, 많은 Labeling 비용 필요
  - 많은 데이터는 거의 항상(Almost always) 성능이 좋아진다.
- 2. 많은 데이터 → 높은 표현력 → 더 좋은 성능
- 3. 하지만, 많은 데이터 → 많은 Labeling 비용
- 4. 사람이 모든 라벨링을 진행하는 것은 조금 아깝다는 생각이 든다.
- 5. 어떤 데이터가 필요한지를 기계가 판단하여 사람에게 라벨링을 부탁한다면
  - 사람은 더 적은 라벨링 공수를 들이고도 좋은 모델을 학습할 수 있지 않을까?

참고 : https://kmhana.tistory.com/4

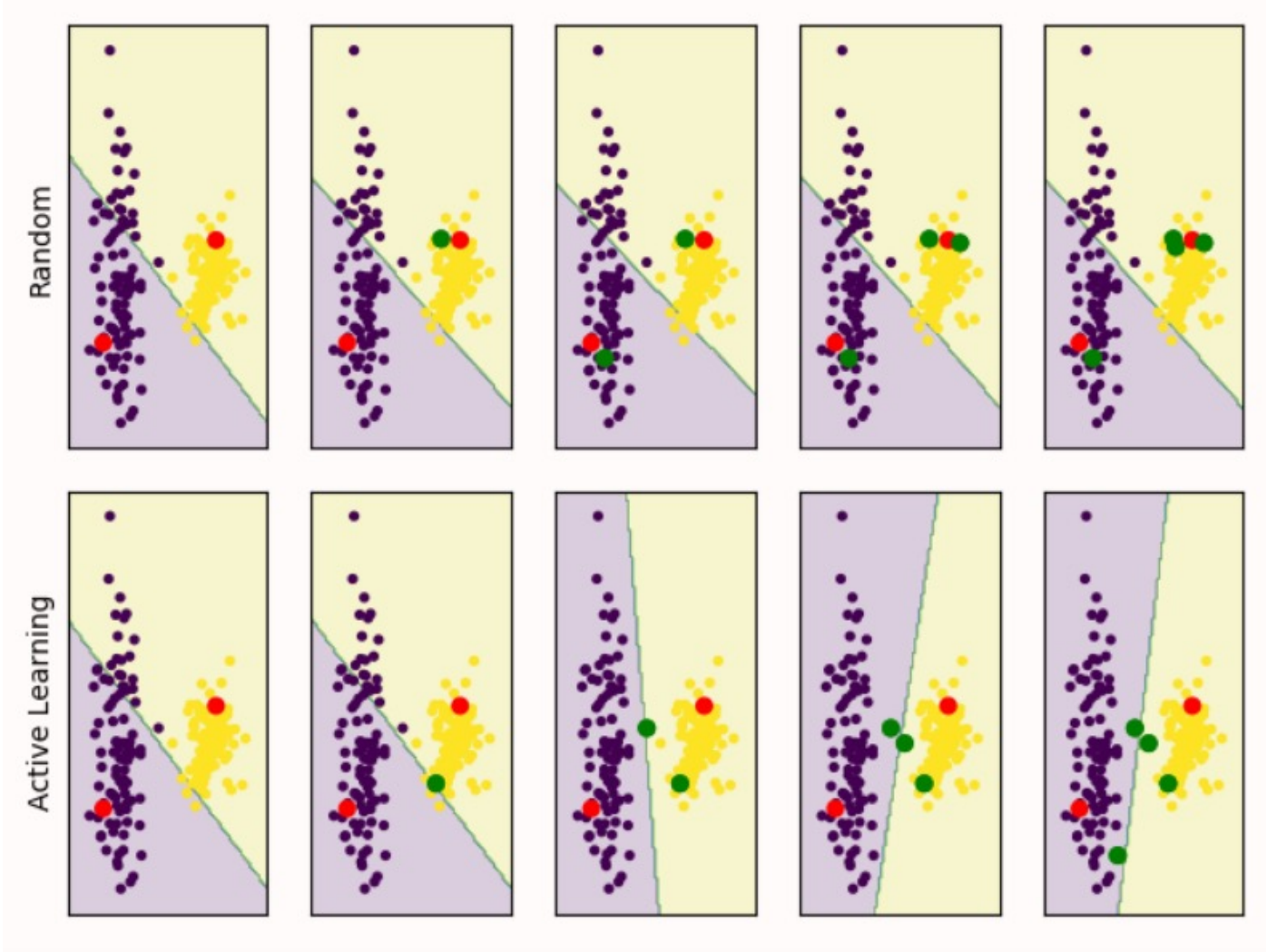
참고 – Active Learning

* 빨간점 : labeling 된 데이터
* 초록점 : labeling을 위해 선택된 데이터

“ 숲길을 따라가며, 나무를 선별하는 방식

혹은

전체 숲을 보며, 중요한 나무를 찾는 방식 ”



참고 : https://kmhana.tistory.com/4

Annotation - 해결방안

- 3. 기존의 구조화 데이터(Existing Structured Data)를 label로 사용
  - e.g. 입원, 사망, 재입원, 응급실 방문 등의 데이터 사용
  - e.g. 과거 데이터에서 예측 모델을 훈련하여
    위험에 처한 환자를 식별하기 위해 ICD 진단 코드 사용
  - 일부 다른 영역에서의 유용성 불분명

- 4. 탐욕적 일치 기반 반자동(semiautomated) 레이블링 [33, 105]

- 5. 크라우드소싱
  - 프라이버시 제약

Provenance (출처)

- 임상 내러티브의 구조와 스타일은 기관마다 크게 다를 수 있음[119]

- 데이터 출처가 소수 기관에 국한되는 경우,
  데이터의 대표성 부족 → 과적합 (overfitting)

- 본 논문에서 검토된 대부분의 연구 데이터는 연구 저자와 관련된 기관으로 한정됨
  - 데이터셋에 자유롭게 접근하기는 어렵기 때문일 것

Types of Narratives

- 대부분 단일 유형의 임상 내러티브에 초점을 맞춤
  - e.g. 심장초음파 보고서는 심혈관 의학과 관련된 정보를 추출하는 데 사용
  - e.g. 뇌파 검사 보고서는 간질을 연구하는 데 사용
  - e.g. 정신과 기록은 건강 정보와 증상 심각도를 추출하는 데 사용 등

Clinical Applications

- 대다수의 연구는 지도 학습에 적합한 텍스트 분류 작업을 수행
  - 분류 모델은 phenotyping(표현형, 형질형), prognosis(예후), 치료 개선, 자원 관리 및 감시를 하는 데 사용

	Classification	Clustering	Coreference resolution	Information extraction	Named entity recognition	Ranking	Word sense disambiguation	**Total**
Care improvement	8			2				**10**
Comparative effectiveness	1							**1**
Data management	2			1	1			**4**
Diagnosis	2							**2**
Efficiency	1							**1**
Enabling	7		2	4	14		3	**30**
Interactive NLP	1			1				**2**
Knowledge acquisition				1				**1**
Patient literacy						1		**1**
Pharmacovigilance	3			1				**4**
Phenotyping	13							**13**
Prognosis	13			3				**16**
Quality	2							**2**
Referral	1							**1**
Resource management	8							**8**
Risk prediction	1							**1**
Safety	4							**4**
Service improvement	1							**1**
Surveillance	5			1	1			**7**
Triage	1	2		1	1			**5**
Unclear	1							**1**
Total	**75**	**2**	**2**	**15**	**17**	**1**	**3**	

Summary - 기계 학습에 사용되는 데이터의 주요 속성 조사

- 훈련 데이터셋의 크기가 상대적으로 작은 경향이 있음을 발견
  - 훨씬 더 큰 데이터셋을 사용할 수 있음에도, 상대적으로 적은 비율만 사용
- 데이터셋은 대부분 소수의 기관에서만 제공받음
- 가장 일반적으로 사용되는 데이터 소스는 MIMIC와 VHA
- 영상 보고서부터 퇴원 요약까지 다양한 종류의 단일 유형의 임상 내러티브에 초점을 맞춤
- 대부분의 학습 데이터는 텍스트 분류, 정보추출 및 개체명 인식에 사용
- 일반적으로 텍스트 분류는 표현형, 예후, 치료 개선, 자원 관리 및 감시와 같은 임상 적용에 사용됨
  - 나머지 NLP 작업에는 명확한 임상 적용이 없었음

IV. DISCUSSION

텍스트 데이터 사용 시 문제

1. 환자의 개인 정보 보호 문제

2. 주석 병목 현상

3. 데이터를 여러 기관에서 얻기 어려움
   → 따라서, 대부분은 한 기관의 데이터만을 사용함
   → 임상 기록의 형식과 스타일은 기관에 따라 다양함
   → 데이터 의 편향성 (대상 문제의 특성 분포를 반영X)
   → 과적합 발생

해결 방안 1, 2, 3

- 1. 데이터 증강 (Data augmentation) 기법 사용
  - 새로운 데이터를 수집하지 않고, 모델의 학습에 사용할 수 있는 데이터를 다양화하고 텍스트 데이터를 보강
- 2. 전이 학습(transfer learning) 적용
- 3. 원격 지도 (distant supervision) 개념을 적용
  - 기존의 구조화 데이터에 의존하여 텍스트 데이터에 자동으로 주석처리
  - 수동 데이터 주석을 완전히 피할 수 있음

해결방안 4

- 4. 지도 학습과 비지도 학습 중 데이터에 적합한 방법론을 선택
  - 지도 학습 접근 방식에 적합 : 레이블을 쉽게 사용할 수 있는 경우
    - e.g. 병원 내 사망, 재입원, 응급실 방문 등
  - 데이터에 처음부터 수동으로 주석을 달아야 하는데 무조건 지도 학습을 주장하는 것은 '둥근 구멍을 통해 사각형 못을 맞추려는 것'과 매우 유사
  - 비지도 학습 방식이(e.g. 토픽 모델링 등) 더 적합할 수 있음에도 불구하고 시도조차 하지 않음

요약

- 데이터 주석 병목 현상을 임상 NLP에서 기계 학습 접근 방식의 주요 장애물 중 하나로 식별
- 능동 학습은 주석 작업을 보다 전략적인 방식으로 접근하는 방법
- 데이터 증강, 전이 학습 및 원격 감독과 같은 대안을 사용하여 이점을 얻을 수 있음
- 궁극적으로 비지도 학습은 데이터 주석의 필요성을 완전히 배제하므로, 임상 NLP에 더 자주 사용해야 함