# 나의 머신러닝 프로젝트의 bias는 무엇이 있을까?

휴먼지능정보공학과 201910803 박채희

기업 이미지 분석을 위해, 기업 이름이 들어간 뉴스, 커뮤니티, sns 글의 데이터들을 크롤링 하여 감성사전을 구축하는 프로젝트를 한 경험이 있습니다. 그리고 뉴스 데이터들을 먼저 분석했을 때는 대체적으로 기업 이미지의 긍정적인 내용이 많이 나왔고 커뮤니티와 sns글도 긍정적인 내용이 많을 것이다 예상을 하여 감성사전 구축은 긍정적인 내용 밖내 없을 것이라고 걱정을 했습니다. 하지만 커뮤니티와 sns의 글을 분석해 본 결과 부정적인 내용이 굉장히 많았고, 결론적으로는 긍정, 부정 비율이 적절하게 나와 알맞은 감성사전을 구축할 수 있었습니다.

이 경험을 통해, 데이터 수집의 모집단을 결정할 때부터 존재하는 편향인 representation bias가 존재할 수도 있다고 생각을 했습니다. 그래서 데이터를 수집하고 분석하는 과정에서 편향이 일어나지 않게 적절한 비율로 데이터를 수집하는 방법이 필요하다고 느꼈습니다.