|  |
| --- |
| 1. **주제 (10점)**   가짜리뷰 판별 프로그램 개발 제안  **(나)반, 15팀, 20213610 최예림** |

|  |  |
| --- | --- |
| **2. 요약 (10점)**  - 소비자들에게 신뢰할 수 있는 정보를 제공하여 합리적인 구매 결정을 하도록 도운다.  - 가짜 리뷰 생성 패턴 분석  - 최신 자연어 처리 기술과 딥러닝 기법을 활용하여 높은 정확도로 가짜 리뷰를 판별  - 새로운 유형의 가짜 리뷰에도 효과적으로 대응 가능 | **3. 대표 그림 (1개 이상, 10점)** |

|  |
| --- |
| **4. 서론 (1장 이내)**  온라인 쇼핑과 배달 서비스의 급증으로 제품이나 서비스에 대한 리뷰의 영향력이 크게 증가하고 있습니다. 하지만 일부 업체들이 마케팅의 일환으로 협찬,광고성 리뷰를 작성하게 하거나 실제 고객 리뷰를 조작하는 등 리뷰의 신뢰성을 훼손하는 사례가 빈번히 발생하고 있습니다. 최근 한 조사에 따르면 전자상거래 사이트의 리뷰 중 약 30%가 가짜일 것으로 추정되고 있습니다. 미국과 영국 등 일부 국가에서는 가짜 리뷰 작성자에 대한 과태료 부과 등의 제재 조치를 시행 중일 만큼 가짜 리뷰에 대한 문제가 심각하다는 것을 알 수 있습니다. 가짜 리뷰로 인해 전체적인 리뷰의 신뢰도가 떨어지고 있어 소비자들이 상품이나 서비스에 대한 정확한 정보를 얻기 힘들어지기 때문에 합리적인 구매 결정을 내리기 어려워지며, 나아가 기업들에게도 부정적인 영향을 끼치게 됩니다.  AI 모델을 이용한 가짜리뷰 작성 또한 문제가 되고 있기에 전통적인 기계학습 기법에 의존하는 기존 프로그램만으로는 가짜리뷰를 선별해내는 것에 한계가 있을 것입니다. 따라서, 최신 자연어 AI 모델과 딥러닝 기술을 도입하여 만든 가짜리뷰 선별 프로그램은 텍스트의 의미와 문맥을 효과적으로 파악하고 가짜 리뷰 를 더욱 높은 정확도로 선별해 낼 수 있을 것이며, 지속적인 모델 개선을 통해 새로운 유형의 가짜 리뷰에도 효과적으로 대응할 수 있을 것으로 예측됩니다. 현재 생성된 가짜 구매후기 탐지에 대한 연구는 대부분 영어에 한정되어 있어 한국어 생성 텍스트에 대한 탐지를 가능하게 한다는 점이 이 프로그램의 차별성이라고 볼 수 있습니다. |

|  |
| --- |
| 5. 본론 (1장 이내)    가짜 리뷰 판별 프로그램은 자연어 처리와 기계학습 기술을 활용하여 리뷰 데이터에서 가짜 리뷰를 자동으로 식별하고 제거합니다. 시스템의 전체 구조는 데이터 수집 및 전처리, 기계학습 모델 학습, 그리고 모델 평가 및 개선의 세 단계로 구성됩니다.  먼저 데이터 수집 단계에서는 온라인 쇼핑몰, 소셜 미디어 등 다양한 경로에서 리뷰 데이터를 수집하고 정제합니다. 이때 리뷰 본문 외에도 작성자, 날짜, 평점 등 실제 사용자인지 검증하는 데에 도움이 될 만한 정보도 같이 수집합니다.이를 통해 비정상적인 행동패턴을 보이는 지 여부도 함께 관측할 수 있습니다. 예를 들어 극단적인 별점 데이터를 주기적으로 보이는 사용자는 정상적인 실사용자가 아님을 감지해 사용자에게 신뢰도가 낮은 리뷰라고 알릴 수 있도록 하는 것입니다.  다음으로 전처리된 데이터를 활용하여 기계학습 모델을 학습시킵니다. 이 과정에서 최신 자연어 처리 기술인 BERT, GPT 등의 사전 학습 언어 모델과 딥러닝 기법을 적용하여 리뷰 텍스트의 의미와 문맥을 효과적으로 파악하고 가짜 리뷰 여부 및 패턴을 파악합니다. BERT는 양방향 트랜스포머 기반의 모델로, 리뷰 텍스트의 문맥을 효과적으로 인식할 수 있기 때문에 가짜 리뷰와 진짜 리뷰의 미묘한 차이를 포착하는 것이 가능하므로 단순히 ‘협찬’, ‘광고’ 라는 단어로만 판단하는 것이 아니고, AI가 만들어낸 리뷰와 같이 새로운 가짜리뷰를 방지할 수 있는 프로그램을 구현하는 데에 기계학습 모델로 사용하는 것이 적합하다고 판단하였습니다.  학습 과정에서는 검증셋을 활용하여 모델의 성능을 주기적으로 평가하여 과적합 여부를 판단하고 하이퍼파라미터를 조정합니다. 마지막으로 테스트셋을 이용하여 최종 모델의 성능을 평가하고 실제 서비스에 배포하는 것까지가 개발 목표입니다.  학습된 가짜 리뷰 판별 모델의 성능은 정확도, 정밀도, 재현율 등의 평가 지표를 통해 종합적으로 검증됩니다. 특히 리뷰 데이터의 불균형을 고려하여 F1 스코어와 같은 평가 지표의 활용이 필요하다고 생각하고, 모델의 예측 결과를 분석하여 모델을 지속적으로 업데이트 해서 새로운 유형의 가짜리뷰에도 효과적으로 대응할 수 있도록 개발해 나갈 필요가 있습니다. |

|  |
| --- |
| **6. 결론**  이러한 프로그램은 사용자의 후기가 필요한 어떠한 사이트, 어플에든 활용 될 수 있으므로 사용성이 뛰어나다고 생각합니다. 후기 옆에 [협찬], [광고], [신뢰도 00%] 등의 정보를 기재해 소비자가 안심하고 정보를 받아들일 수 있도록 하여 현명한 소비가 가능하게 만들 수 있습니다. 협찬, 광고가 무조건 나쁜 것은 아니지만, 소비자가 광고임을 인지하고 구매한 것과, 실제 후기인줄 알고 속아서 구매한 것은 엄연히 다르기 때문에 소비자의 알 권리를 지키는 프로그램이 필요하다고 생각하였습니다. 기업 입장에서도 공정한 경쟁 환경이 조성되어 윤리경영을 실천할 수 있게 될 것이므로 소비자와 기업 모두에게 긍정적인 영향을 줄 것으로 예상됩니다.  가짜 리뷰를 판별하는 것에서 나아가 생성 기법까지 탐구해야 할 것입니다. 최근 GPT-3 등 대규모 언어 모델의 발전으로 인해 자연스러운 텍스트 생성이 가능해졌기 때문에 이를 활용하여 가짜 리뷰 생성 패턴을 분석하고, 나아가 가짜 리뷰를 사전에 차단할 수 있는 기술을 개발할 수 있고, 이를 통해 가짜 리뷰 문제를 근본적으로 해결할 수 있을 것이라고 생각합니다.  그러나, 가짜 리뷰의 유형과 수법이 지속적으로 진화하고 있어 일률적인 패턴을 포착하기 힘들고 일부 리뷰의 경우 진짜인지 가짜인지 판단하기가 모호하다는 문제가 존재합니다. 만약, 제품의 장단점을 균형 있게 서술했다면 이를 진솔한 리뷰로 볼 것인지, 아니면 가짜 리뷰로 의심할 것인지 판단하기 어려울 것입니다. 이처럼 리뷰의 특성상 사람의 주관적 판단이 개입되므로 현재의 프로그램에는 한계가 있습니다. 따라서 가짜 리뷰 판별 과정에서 인간의 개입과 검증을 일부 수용하는 과정이 필요하고, 가짜 리뷰 작성자들이 프로그램의 탐지 방식을 역공학하여 새로운 우회 기법을 개발할 가능성도 배제할 수 없으므로 이에 대응하기 위한 지속적인 모델 개선과 함께 새로운 데이터 수집 및 전략을 추가해 보안해야 한다고 생각합니다. |

**7. 출처**

[1] 리뷰이미지 출처: <https://www.danbinews.com/news/articleView.html?idxno=21976>