"안녕하십니까, 4조 '진실의 딥' 발표를 시작하겠습니다. 저는 발표를 맡은 구경서입니다. 오늘 저희 팀은 AI가 작성한 리뷰와 인간이 작성한 리뷰를 구분하는 딥러닝 프로젝트에 대해 이야기를 나누려고 합니다. 리뷰 속 숨겨진

진실을 파헤쳐보는 재미있는 여정이 될 것입니다. 함께 해주세요!

1 발표순서는 다음과 같습니다.

2 팀원역할과 수행과정은 이렇게 다들 한땀한땀 고생하셨습니다.

3. 그럼 첫번째 주제 선정 배경부터 말씀을 드리겠습니다.

4. 먼저, 리뷰의 중요성에 대해 먼저 말씀드리겠습니다. 소비자들은 제품의 품질이나 성능을 직접 확인할 수 없기 때문에, 리뷰와 평점은 구매 결정에 있어 핵심적인 기준이 되고 있습니다. 한 설문 조사 결과, 다른 요소보다 리뷰가 구매에 미치는 영향이 매우 크다는 것을 볼 수 있었습니다.

오른쪽에 보시면 소비자 리뷰에 대한 전반적인 인식을 보았을 때도 구매 의사결정에서 리뷰가 중요한 역할을 함을 알 수 있었습니다.

5 다음으로 지금 보시는 뉴스들과 같이 오늘날의 AI 기술은 사람처럼 자연스럽고 설득력 있는 리뷰를 생성할 수 있어, 소비자들이 가짜 리뷰로 인해 피해를 보는 사례가 많아지고 있습니다.

일부 기업은 고의적으로 긍정적인 AI 리뷰를 대량 생성하거나, 경쟁사를 깎아내리기 위해 부정적인 허위 리뷰를 작성하기도 하고 있습니다. 실제로 평점 후기 알바를 모집하는 사례도 있으며, 알바는 특정 제품의 평점을 조작하거나 허위 리뷰를 작성하도록 유도하고 있습니다.

6. 자 이제 본격적으로 시작하겠습니다. 먼저, 이번 프로젝트의 목표를 말씀드리겠습니다. 저희의 목표는 AI가 작성한 리뷰와 인간이 작성한 리뷰를 구분할 수 있는 모델을 생성하는 것입니다.

7. 그럼 데이터 설명부터 해보겠습니다.

8. 저희가 사용한 데이터를 설명드리겠습니다. 이번 프로젝트에서는 영어 리뷰 데이터와 한국어 리뷰 데이터를 활용하였으며, 먼저 영어 리뷰 데이터를 설명드리겠습니다. 영어 데이터는 총 40,432행 4열로 구성되어 있습니다.

리뷰는 AI가 작성한 리뷰 20,000개와 인간이 작성한 실제 제품 리뷰 20,000개로 이루어져 있습니다. 컬럼을 설명 드리자면 카테고리는 총 10가지로 구분되어 있으며 그리고 rating은 리뷰 평점을 나타내는 값으로 1점부터 5점까지 분포되어 있습니다.

또한, label은 리뷰 작성 주체를 구분하는 항목이며, CG는 AI가 작성한 리뷰, OR은 인간이 작성한 리뷰를 의미합니다. 마지막으로, text는 리뷰의내용을 포함하고 있습니다.

9. 다음으로 한글 데이터에 대해 설명드리겠습니다. 저희는 네이버 스마트스토어에서 리뷰 데이터를 수집하기 위해 7가지 카테고리를 선정하여 데이터를 크롤링했습니다. 이를 통해 약 27000개의 리뷰 텍스트를 확보하였습니다.

수집된 데이터는 중복 리뷰, 재구매 리뷰, 그리고 한달 사용 리뷰를 제거하고, 챗지피티를 통해 AI의 리뷰를 1:1 비율로 생성하여 최종적으로 약 54,000개의 한글 리뷰 데이터를 구성하였습니다.

10 다음으로 EDA과정을 말씀드리겠습니다.

11. 영어 리뷰데이터에서 AI와 인간의 리뷰 단어 빈도수를 막대그래프로 시각화 해보았습니다

AI 리뷰에서 가장 많이 사용된 단어는 'loves', 'great', 'son'으로, 감정적으로 강한 긍정적 표현과 특정 대상에 집중된 단어가 주를 이루고 있습니다.

반면, 인간 리뷰에서는 'great', 'old', 'like'가 상위 단어로, 보다 다양하고 구체적인 단어들이 포함되어 있음을 알 수 있습니다.

이러한 차이는 AI 리뷰가 감정을 강조하는 경향이 강한 반면, 인간 리뷰는 경험과 의견을 더 자연스럽게 표현하는 데 초점이 맞춰져 있다는 점을 보여줍니다."

12 다음으로는 워드클라우드로 AI와 인간 리뷰 단어빈도수를 한글 리뷰 데이터로 분석하여 시각화한 결과입니다.

AI가 작성한 리뷰에서는 '아주', '정말', '덕분에' 같은 과장된 긍정 표현이 두드러지게 나타났습니다. 이는 AI 리뷰가 감정적으로 긍정적인 이미지를 강조하려는 경향이 강하다는 점을 보여줍니다.

반면, 인간이 작성한 리뷰에서는 '배송', '좋아요', '좋습니다'와 같은 구체적이고 현실적인 표현들이 주로 사용되었으며, 실제 경험을 바탕으로 한 서술이 많았습니다.

이를 통해 AI와 인간 리뷰의 차이점을 명확히 시각적으로 확인할 수 있었습니다."

13 다음으로 영어 리뷰의 품사 비율을 분석해본 결과로 전치사, 부사, 고유명사에서 눈에 띄는 차이가 발견되었습니다. 특히 주목할 만한 점은 고유명사의 사용 빈도로, 인간이 작성한 리뷰에서 ai생성 리뷰보다 약 3배 높은 비율을 보였습니다. 전치사와 부사의 높은 사용 빈도를 보아 인간 작성 리뷰가 더 복잡하고 정교한 문장 구조를 가지고 있음을 시사하고, 이를 통해 인간이 특정 제품을 더 다양한 문장으로 리뷰함을 알 수 있었습니다

14 마찬가지로 한국어 리뷰도 품사 비율 분석 결과가 인간이 ai보다 명사, 동사, 조사를 훨씬 풍부하게 활용하여 자연스럽고 구체적인 문장을 구성한다고 볼 수 있습니다. 눈에 띄는 점은 ai가 작성한 리뷰는 문장부호를 다수 포함하며 형식적이고 단순한 구성을 보인다는 것입니다. AI 생성 리뷰는 아직 인간의 자연스러운 언어 사용을 완벽히 모방하지 못하고 있음을 보여줍니다.

15 AI가 생성한 리뷰와 인간이 작성한 리뷰에서 문법적 오류 및 어색한 표현의 유무를 분석하기 위해 기준을 설정하고 이를 비율로 나타내어 비교하였습니다.

저희가 정의한 오류 판단 기준은 다음과 같습니다

이 기준에 따라 분석한 결과 AI가 인간보다 문법 오류 비중이 높음을 알 수 있었습니다

이 결과는 AI가 리뷰를 생성하는 과정에서 문법적 오류를 빈번히 일으킨다는 점을 보여줍니다. 특히, AI는 단어를 생성할 때 앞뒤 단어와의 관련성을 기반으로 선택하는 과정에서 문법 규칙을 충분히 고려하지 못한다는 한계를 드러냈습니다.

16 다음으로 "코사인 유사도 분석을 통해 AI 리뷰와 인간 리뷰 간의 평균 유사도를 비교해본 결과, AI 리뷰가 인간 리뷰보다 평균적으로 높은 유사도를 보였습니다. 이는 AI가 리뷰를 생성할 때, 기존에 학습된 데이터를 기반으로 유사한 패턴을 반복적으로 생성하는 경향이 있기 때문입니다. 즉, AI는 특정 키워드나 표현을 자주 사용한 반면, 인간이 작성한 리뷰는 개인의 표현 방식이나 주관적인 의견을 나타냅니다. 이와 같은 분석을 통해 AI와 인간 리뷰의 작성 스타일 차이를 명확하게 구분할 수 있었습니다."

17 "다음으로, 감성 분석을 진행하였습니다. Sentiment Intensity Analyzer를 활용하여

다음과 같은 각 기준을 통해 리뷰를 긍정, 중립, 부정으로 분류를 해보았고 이것을 통해 리뷰 유형별 감성 점수와 평점 간의 상관관계를 알아보았습니다. 분석 결과, AI가 작성한 리뷰에서 평점과 감성 간의 상관계수가 낮아, 감정이 평점과 일치하지 않는 경우가 많다는 것을 볼 수 있었습니다. 이러한 특성은 가짜 리뷰 탐지 모델의 필요성과 유용성을 뒷받침하는 자료로 사용할 수 있었습니다.

18 네 다음으로는 저희 여혁수팀장님이 발표를 이어서 하겠습니다.

– 데이터 전처리 파트부터는 제가 이어가겠습니다.

19 – 우선 영문 리뷰 데이터셋의 경우 일부 데이터가 중간에 텍스트가 잘려 있었습니다. 온라인 쇼핑몰에 리뷰를 보다보면 길어서 중간에 잘리고 더보기 버튼이 있는 경우를 한번씩 보셨을 거라 생각하는데요, 그런 케이스 때문에 저희가 보유한 데이터셋에도 텍스트가 잘려 나왔다고 추정을 합니다. 그래서 저희는 리뷰의 잘린 부분을 완성하는 모델을 만들어서 텍스트를 채웠습니다. 잘린 리뷰라고 판단한 기준은 우선 마침표나 물음표, 느낌표로 끝나지 않는 것을 확인하구요, 그리고 저희 모델이 해당 리뷰의 바로 뒤에 올 단어로 end token을 상위 10% 내로 높은 확률로 예측되지 않는 것으로 정했습니다.

모델은 CNN과 LSTM이 결합된 구조와 단순 LSTM으로 실험을 해봤고, 정확도가 잘 나오는 단순 LSTM 모델을 적용해서 오른쪽 예시와 같이 텍스트를 생성했습니다.

20 – 다음으로는 한글 리뷰 전처리입니다. 한글 데이터는 형태소 분석과 어근 추출이 필수적이었습니다. 예를 들어 조용히, 조용하게, 조용한 과 같이 한글은 의미는 같지만 형태는 다양한 경우가 많습니다. 이러한 단어를 분해해서 의미있는 부분만을 추출하는 데 바로 이 konlpy 라이브러리가 사용됩니다. Konlpy 내에도 다양한 분석기가 있는데요, 저희는 리뷰 텍스트 특성상 띄어쓰기 오류에도 내구성이 강한 분석기가 필요했기 때문에 꼬꼬마를 사용했습니다.

21 – konlpy 적용 예시인데요. 아래와 같이 문장을 분해하게 됩니다.

22 – 이제 모델링으로 넘어가보도록 하겠습니다. 먼저 영문 리뷰를 구분하는 모델입니다.

23 – 먼저 가장 기초적인 딥러닝 모델 MLP입니다. 우선 TF-IDF라는 벡터화 방식을 통해서 텍스트 데이터는 단어의 등장 빈도수가 나열된 수치 데이터로 변환됩니다. 변환된 입력데이터는 은닉층으로 들어가서 패턴을 학습하는 데에 사용되고, 마지막 sigmoid를 통해 AI와 인간으로 이진 분류가 됩니다.

24 – 다음은 이미지 데이터 학습에 많이 사용되는 CNN입니다. CNN이 텍스트 데이터와 만나면 문맥적인 특징 추출이 가능합니다. 우선 MLP와 마찬가지로 데이터의 벡터화가 필요하구요, 컨볼루션 레이어와 맥스풀링을 거쳐서 특징들이 추출이 됩니다. 마지막으로 MLP와 동일하게 sigmoid 거쳐서 이진 분류가 됩니다. 이러한 CNN으로는 문맥을 학습할 수는 있지만, 순차적 정보를 학습하는 것은 무리가 있습니다. 그래서 다음으로 소개해드릴

25 – LSTM은 CNN의 단점을 보완하는 데에 아주 적합한 모델이라 할 수 있습니다.

LSTM은 데이터에 있는 정보가 수많은 LSTM 유닛들을 거치면서 중요한 정보는 저장하고 불필요한 정보는 망각하면서 순서에 따른 의존 관계를 학습할 수 있습니다.

26 – 다음은 3가지 모델에 대한 평가입니다. 평가 기준으로는 정확도, 모델 복잡도, step당 학습시간 이렇게 3개를 선정했습니다. CNN이 가장 정확도가 높았고, 모델을 구성하는 매개변수는 제일 많았지만 학습시간은 LSTM보다 효율적이었습니다. 이 결과를 통해서 CNN도 텍스트 분류 문제라면 충분히 활용성이 높다는 것을 알 수 있었습니다.

27 – 다음은 한글 리뷰를 구분하는 모델로 넘어가보겠습니다.

28 – 영문 리뷰 모델링에 사용된 것과 동일한 모델을 적용했기 때문에 중복적인 부분은 제외하고 설명 드리겠습니다. 일단 영문 리뷰와는 다르게 띄어쓰기 보정과 형태소, 어근 추출이 추가적으로 필요했기 때문에, PYKOSPACING, Konlpy로 전처리 과정을 거쳤습니다. MLP의 정확도는 99.38%로 AI, 인간 구분없이 균형있게 잘 맞췄습니다.

29 – 다음으로 CNN은 정확도 99.71%로 확실히 MLP보다 한 단계 더 좋은 성능을 보였습니다. 영문 뿐만 아니라 한글 데이터에서도 앞뒤 문맥을 좁게 추출해서 학습하는 것이 효과가 좋았다고 볼 수 있겠습니다.

30 – 다음으로 LSTM입니다. 정확도 99.72%로 현재까지 가장 좋은 성능을 보였지만 CNN과는 0.01% 차이로 매우 근소했습니다. 영문과 한글 리뷰에 대한 모델을 전체적으로 봤을 때 CNN이 효율이나 성능 어떤 면에서 보더라도 우수한 것을 보고 저희는 이 CNN의 국소적인 패턴 학습이 충분히 강력해서 순차적 정보까지도 캐치를 한게 아니냐는 생각을 하게 되었습니다.

31 – 마지막으로 KoGPT2라는 모델인데요. 저희가 아직 배우지 않은 자연어처리 모델이지만 한글 리뷰 데이터에 한해서는 한국어 처리에 특화된 모델을 한번 경험해보자는 의견이 모아져서 도전하게 되었습니다. KoGPT2는 40기가 이상의 대규모 한글 데이터를 받아들여서 사전에 학습이 되어 있는 언어모델이구요. 주로 챗봇 응답 생성, 그리고 문서 요약에 사용되는 모델이라고 합니다. 모델의 세부구조는 오른쪽의 다이어그램과 같습니다. 참고하시면 되겠습니다.

32 – kogpt는 사전 학습된 모델이기 때문에 내부 구조가 공개되어 있지 않습니다. 매개변수는 1억개가 넘었구요. 설정할 수 있는 하이퍼파라미터도 그렇게 많지 않았습니다. 학습율과 배치 사이즈를 조절해서 학습의 속도와 안정성을 둘 다 잡으려고 했구요. 그리고 사용할 수 있는 메모리를 최대한으로 활용할 수 있는 선에서 입력 데이터의 최대 길이를 결정했습니다. 정확도는 99.16%로 기대보다 낮았습니다. 단순 분류 문제에 이렇게 고도화된 모델은 과하고, 확실히 분류보다는 생성에 최적화된 모델임을 알 수 있었습니다.

33 - 지금까지 소개해드린 4가지 모델의 성능을 표로 정리해봤습니다. 4개 모두 정확도가 99퍼를 넘겼는데, 이거는 저희가 임의로 생성한 데이터가 많다보니까 AI와 인간의 패턴차이를 학습하기가 쉬워진 게 아닌가 생각을 합니다. 정확도는 LSTM이 근소하게 1위를 차지했습니다만, 모델 구조의 복잡도, 학습 효율까지 전부 고려한다면 CNN이 이 중에서는 가장 합리적인 선택이라고 할 수 있겠습니다.

34 - 다음으로 실제 리뷰에 적용한 예시입니다. 정확도가 가장 좋았던 모델로 학습 데이터에 사용되지 않은 제품에 대한 리뷰를 긁어와서 얼마나 높은 확률로 분류를 하는지를 나타냈습니다. 저의 최애 소스에 대한 리뷰를 가져왔구요. 쿠팡의 리뷰 중에서는 99.76%로 강하게 AI라고 예측한

리뷰가 있었습니다. 그래서 저의 팀원이 직접 리뷰를 작성했더니 다행히도 정반대로 99.97%로 인간이라고 판단을 했습니다.

35 - 네이버에서도 동일 제품에 대한 리뷰를 긁어왔는데 전부 사람이 쓴 리뷰라고 강하게 예측되었습니다. 네이버는 작년 2월부터 리뷰 클렌징 시스템을 운영하고 있어서, AI로 예측되는 리뷰는 찾기 힘들었던 것 같습니다.

36 - 마지막 파트 결론입니다.

37 - 기대효과는 크게 4가지로 나눠보았습니다. 일단 소비자 입장에서는 AI리뷰로 의심되는 리뷰는 안보이게 한다거나, 의심마크를 붙이거나 해서 합리적인 구매를 돕고, 피해를 최소화할 수 있습니다. 기업의 입장에서는 저희가 만든 모델링 시스템을 통해서 소비자와의 신뢰를 쌓고, 기업들 간의 투명한 경쟁을 촉진할 수 있습니다.

38 - 한계점으로는 우선 전처리 파트에서 말씀드렸듯이 텍스트를 임의로 복원하고, 생성했던 부분이 있어서 데이터의 완전성이 보장되지 않았다는 한계가 있고,

네이버 스마트스토어의 한글 리뷰는 클렌징 시스템을 통과했기 때문에 100% 사람이 작성한 것으로 가정을 했지만 AI 리뷰일 가능성을 배제할 수는 없었습니다. 또한 지금에도 발전중인 AI에 따라서 생성되는 패턴은 저희 모델이 반영하지 못할 수 있고, 많은 모델을 실험하다보니 최적화가 충분하지 못했습니다.

39 - 네 땡스 들어주셔서 감사합니다.