대본

P\_1

안녕하십니까. 항공대 석사 1학기 이성우입니다. 딥러닝 기반 텍스트 분류 모델을 활용한 국내 이커머스 사용자 리뷰 분석발표 시작하겠습니다.

P\_2

연구의 배경부터 말씀드리겠습니다. 다들 아시다시피 디지털 전환에 의한 뉴노멀은 산업의 트렌드로 자리 잡았고 유통업도 예외는 아닙니다. COVID 19으로 인한 언택트 소비증가와 1-2인가구 증가로 인해 고객의 요구는 다양해졌고, 이에 따라 기존의 유통업 중심의 대기업들도 온라인 사업을 본격화하기 시작했습니다.

P\_3

온라인 쇼핑의 거래액은 20년기준 전년대비 19%성장하였고, 전체 소매판매액 중 온라인 쇼핑의 비중은 27%로 집계되었습니다.

본 연구에서는 개인화된 물류 서비스의 필요성을 인식하여, 국내 대표 이커머스 업체인 쿠팡과 후발주자인 롯데ON, SSG닷컴의 사용자 리뷰를 분석하고 고객만족을 분석하겠습니다.

P\_4

참고문헌입니다. 참고문헌은 알고리즘 개선보다는 시사점 도출에 초점을 맞춘 논문을 선정하였습니다. 해당 논문에서는 본 연구와 비슷하게 헬스 앱 사용자의 리뷰를 4가지 컨텍스트로 나누어 향후 디자인을 전략을 제시하였습니다.

P\_5

두번째 논문도 마찬가지로 부동산 앱에서 거주자의 공간에 대한 평가를 수집하여 감성분류모델을 구성하였습니다.

P\_6

전체적인 Frame Work는 다음과 같습니다. 모델링을 위한 데이터를 수집하여 Labeling, Cleaning, Tokenization을 거쳐 감정분류기를 구성합니다. 분석을 위한 데이터를 수집하여 Cleaning, Tokenization을 거친 후, 감정분류기를 거쳐 Sentiment를 부여하고, 토픽 사전을 통해 토픽을 부여합니다. 완성된 Data를 다방면으로 분석하면서 연구가 마무리됩니다.

P\_7

첫번째로 Corpus수집입니다. Python의 Selenium을 이용하였고 Modeling을 위한 데이터의 경우 쿠팡이나 네이버 등의 쇼핑몰에서 약 19만개의 데이터를 수집하였습니다.

분석 데이터의 경우 구글 플레이스토어와 애플 앱스토에서 약 8만개의 데이터를 수집하였습니다.

Labeling의 경우 Modeling Data는 별점을 기준으로 5, 4점은 Positive, 2, 1점은 Negative로 부여 후 연구자가 부적절하게 평가된 리뷰는 제거하였고 Class의 비율은 50대 50으로 맞췄습니다. 분석 Data Set의 경우 Topic Classifier의 평가를 위해 2000개의 Review Data의 Topic을 Labeling하여 준비하였습니다.

P\_8

Cleaning은 정규표현식을 이용하여 전각문자와 의미없이 나열된 특수문자를 없애줬고, 영어의 경우 대문자를 소문자로 통일하였습니다.

한국어는 교착어의 특성을 가지고 있기 때문에 토큰화가 매우 중요합니다. 접사를 분리하여 Vocabulary의 사이즈를 줄여서 Sparseness를 줄였습니다.

P\_9

LSTM의 경우 RNN의 장기 의존성 문제 즉, Gradient Vanishing 문제를 해결하기 위해 고안된 알고리즘입니다. NLP에서 널리 쓰이고 있고, 어느정도 문장의 길이가 길어져도 Gradient Vanishing을 해결할 수 있다고 알려져 있습니다.

다음으로 CNN입니다. CNN은 원래 Computer Vision이나 Speech Recognition에서 주로 사용되는 알고리즘으로 데이터의 특징을 잘 파악한다는 특징이 있습니다.

문제 정의입니다. Sentiment Classifier는 Text를 입력받아 Class를 반환합니다. 즉. Review를 입력받고 긍정, 부정 두개의 클래스를 반환합니다.

P\_10

Sigmoid를 Activation Function으로 하는 forget, input, output gate가 존재하여 Gradient Vanishing을 문제를 해결한 LSTM입니다.

P\_11

CNN은 보시는 것처럼 kernel 그러니까 Window가 해당 문장을 몇 개의 임베딩 단위로 매칭하면서 매칭이 되는 특징이 있는지 살피는 방식으로 진행됩니다. 마지막 linear layer를 거치면서 확률 분포를 반환합니다.

P\_12

Hyper Parameter Tuning은 다음과 같이 진행하였습니다. LSTM의 경우 연구자의 로컬 컴퓨팅 환경으로 고려하여 NLP에서 일반적으로 쓰이는 Parameter로 고정하였고, CNN의 Window Size와 filter의 개수만을 고려하였습니다.

30,000개의 Test Data Set으로 3번째 모델의 Accuracy가 가장 정확하여 3번째 모델을 선택했습니다.

해당 모델의 Validation loss와 accuracy는 다음과 같습니다.

P\_13

다음으로 Topic Classifier입니다. 사실 처음에는 딥러닝 기반 멀티클래스 분류기를 구성했었지만 Labeling의 한계로 성능이 매우 좋지 않았습니다. 대안으로 각 토픽에 대한 키워드를 추출하여 그 핵심 키워드의 포함 여부에 대해 리뷰를 분류하는 방법을 사용하였습니다. 연구자가 약 만개정도의 리뷰를 확인하여 8개의 Topic을 지정했고 Topic별로 시드단어를 선정했습니다. KeyWord 확장을 위해 Fasttext를 이용하여 Review Data에서 Seed Word와 높은 유사도를 가진 동시 출현 단어들을 추출하였습니다.

P\_14

2000개의 Test Data로 평가를 해본 결과입니다. 데이터가 매우 편향되어 있어 정확히 예측하는것에는 한계가 있었습니다. 이는 연구의 한계점으로 추후 연구 과제로 남겨두었습니다.