대본

P\_1

안녕하십니까. 항공대 물류과 석사과정 이성우입니다. 딥러닝 기반 텍스트 분류 모델을 활용한 국내 이커머스 사용자 리뷰 분석발표 시작하겠습니다.

P\_2

연구의 배경부터 말씀드리겠습니다. 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.

분석 모델로는 LSTM과 CNN의 앙상블 모델을 구축하여 감성분석을 진행하고, Fasttext를 활용하여 리뷰데이터를 8가지 토픽으로 분류하는 Dict를 만들었습니다.

P\_4

참고문헌입니다. 참고문헌은 알고리즘 개선보다는 시사점 도출에 초점을 맞춘 논문을 선정하였습니다. 해당 논문에서는 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. 감성분류 모델링을 위한 데이터를 수집하여 Labeling, Cleaning, Tokenization을 거쳐 감성분류기를 구성합니다. 분석을 위한 데이터를 수집하여 Cleaning, Tokenization을 거친 후, 감성분류기를 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.

본 연구에 직접적으로 사용되는 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

모델링에는 파이썬의 프레임워크인 Pytorch를 활용하였습니다. 알고리즘은 LSTM과 CNN을 사용했습니다. 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 존재하며 이를 열고 닫으면서 Gradient Vanishing을 문제를 해결한 LSTM입니다.

P\_11

CNN은 보시는 것처럼 kernel 그러니까 Window가 해당 문장을 몇 개의 임베딩 단위, 즉 윈도우 사이즈로 매칭하면서 매칭이 되는 특징이 있는지 살피는 방식으로 학습이 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. 다음과 같은 파라미터로 세가지 모델을 구성했습니다.

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 않았고, 데이터가 매우 편향된 특징을 가지고 있어서 모델링에 어려움이 있었습니다. 대안으로 각 토픽에 대한 키워드를 추출하여 그 핵심 키워드의 포함 여부에 대해 리뷰를 분류하는 방법을 사용하였습니다. 연구자가 약 만개정도의 리뷰를 확인하여 8개의 Topic을 지정했고 Topic을 설명하는 Seed Word를 선정했습니다. Seed Word를 활용하여 KeyWord를 확장을 위해 Fasttext를 이용하였고, Review Data에서 Seed Word와 높은 유사도를 가진 동시 출현 단어들을 추출하여서 KeyWord에 추가하였습니다.

P\_14

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

P\_15

분석 데이터를 모델에 적용하였습니다. 리뷰는 18년 1월 1일부터 21년 5월 19일까지 수집하였고 각각 쿠팡 62,375개, 롯데ON 3,328개, SSG닷컴 8,784개가 수집되었습니다.

죄측 하단은 태깅된 토픽의 수에 따른 리뷰수를 보여주고있습니다. 아무 토픽도 태깅되지 않은, 그러니까 None타입의 리뷰가 제일 많은 것을 볼 수 있구요. 하나의 토픽만을 가지는 리뷰가 27%정도 되는 것을 알 수 있습니다. 우측상단은 각 토픽이 얼마나 많은 리뷰에 태깅되었는지를 보여주는데, 상품, 배송 그리고 앱에 대한 리뷰가 많은 것을 알 수 있구요. 우측 하단은 그렇게 달린 각 토픽별 긍정비율을 나타내고 있습니다. 상품, 배송, 가격의 경우 긍정적 리뷰가 많이 달렸지만 나머지는 그렇지 않은 것을 확인할 수 있습니다. 특히 광고는 긍정리뷰가 거의 없었구요. None타입의 경우 긍정리뷰가 압도적인데, 이는 부정적 리뷰의 경우 짧게 쓰기보단 불만을 길게 나타낸다는 것을 알 수 있습니다.

P\_16

결과 탐색입니다. 롯데ON과 SSG닷컴의 디지털전환 전 후를 비교했고 각 데이터를 IPA방법으로 토픽을 분류했습니다. 마지막으로 쿠팡의 데이터도 같은 분석을 진행했습니다.

각 회사의 리뷰데이터를 디지털전환 전 후로 나누었습니다. SSG닷컴의 경우 디지털 전환 전인 신세계몰과 이마트몰의 리뷰수 약 1500개, 긍정비율 60%, 롯데ON의 전신인 롯데마트의 리뷰수 약 1200개 긍정비율 38%, 롯데ON리뷰수 약 7500개, 긍정비율 54% 마지막으로 쿠팡은 디지털 전환으로 볼 순 없지만, 회사간 비교에서 시기를 맞춰주기 위해 롯데ON의 디지털 전환 후인 20년 4월이후의 데이터를 나눴습니다. 이전 리뷰수 약 53000개 긍정비율 85%, 이후 리뷰수 약 9000개 긍정비율 53%입니다.

P\_17

먼저 롯데ON입니다. 좌측 상단은 디지털 전환 전 각 토픽별 긍정비율을 나타내고 있구요, 좌측 하단은 디지털 전환 후 각 토픽별 긍정비율을 나타내고 있습니다. 우측 상단은 디지털 전환 전 후의 긍정비율에 대한 z-test의 p-value값을 보여줍니다. 디지털 전환 후 Price를 제외한 모든 토픽에서 긍정리뷰가 유의하게 증가한 것을 알 수 있습니다. 우측하단의 IPA매트릭스를 보시면 상품의 경우 1사분면에 위치하여 중요도와 만족도 모두 증가한 것을 볼 수 있습니다. 나머지 토픽들도 전반적으로 우측상단으로 위치하는 것으로 보이는데 디지털 전환이 기업경영에 긍정적 영향을 미쳤다고 해석할 수 있겠습니다. 마지막으로 앱리뷰가 높은 중요도를 가지는 것은, 앱스토어 리뷰라는 데이터의 특성에 기인하는 것이라고 볼 수 있겠습니다.

P\_18

다음으로 SSG닷컴입니다. SSG닷컴은 롯데ON과는 상반되게 디지털 전환 후 긍정비율이 떨어지거나 비율의 차이가 유의하지 않다는 결과를 얻었습니다. 특히 Goods와 배송 결제수단 등에서 퍼포먼스가 많이 떨어졌습니다. 20년 기준 시장 점유율 3%로 SSG보다 디지털 전환을 늦게한 롯데온의 5%보다 떨어지는 결과를 보여주고있습니다. Goods의 경우 좋은 평가를 얻었지만 롯데ON에 비해 배송과 결제에서 좋지 못한 평을 보여주었습니다.

P\_19

마지막으로 쿠팡입니다. 연구 초기에는 시장의 선도기업인 쿠팡의 데이터를 모범데이터로 하여 앞선 두기업과 비교할 생각이었지만, 쿠팡의 데이터를 분석해 본 결과 롯데ON의 디지털 전환 시기인 20년 4월 28일 이후 데이터가 그 전 데이터에 비해 지표가 매우 떨어지는 것을 볼 수 있습니다. 거의 모든 토픽에 대해 유의하게 긍정비율이 떨어졌습니다. 따라서 연구자는 쿠팡의 최근데이터와 다른 기업을 비교하는 것은 무리라고 판단했고, 과거의 데이터 또한 여러가지 요인 특히, COVID 19같은 외적요인을 통제할 수 없다고 판단하여, 쿠팡의 전 후 데이터만을 비교하는 것으로 결론을 냈습니다. 거의 모든 토픽에 대해 긍정비율이 떨어졌지만 특히, 배송과 상품이 많이 하락한 것으로 나타났는데, 리뷰를 분석해본 결과 배송의 경우 노동자 사망 및 혹사 문제가 부정적 영향을 끼친 것으로 나타났고, 상품의 경우 짝퉁 및 중국산 제품을 많이 쓰는 것 같다는 평이 있었습니다.

P\_20

마지막으로 추후 연구 과제를 말씀드리고 마치겠습니다. 먼저 Topic Classifier의 성능이 매우 좋지 않아 많은 아쉬움이 있습니다. 이를 개선하기 위한 연구를 할 계획입니다. 두번째로는 최신 NLP알고리즘을 연구하여 Sentiment Classifier에 적용할 계획입니다. 세번째는, 이번에는 어플리케이션 마켓의 리뷰만을 이용하였지만, 이는 지나치게 편향된 자료이므로 마켓의 리뷰 데이터 외의 블로그 등의 유저의 리뷰를 수집하여 분석할 계획입니다. 마지막으로 본 연구의 Sentment Classifier는 중립 리뷰를 고려하지 않았는데 추후에는 이를 고려하는 멀티클래스 분류기를 연구할 계획입니다. 이상으로 발표를 마치겠습니다. 감사합니다.