Term Project - 9조 Proposal

1. 풀고자 하는 문제 정의

오늘날 대다수의 소비자들은 구매 결정 시 온라인 리뷰를 읽는다. **온라인 리뷰(Online Review)란, 제품이나 서비스를 구매한 후 이를 경험한 소비자들이 기업, 제품, 서비스와 관련된 평가적인 내용을 자신이 이용한 온라인 판매처 또는 SNS 등에 글, 사진, 동영상 등의 형태로 포스팅하는 것을 의미한다.** 최근 소비의 행태가 오프라인에서 온라인 중심으로 변화하며 수많은 온라인 리뷰 데이터들이 누적되고 있지만, 정작 해당 리뷰를 읽는 잠재 고객들은 그 모든 온라인 리뷰들을 소화하기 어려워 할 때가 많다. 왜냐하면 대부분의 리뷰들은 단순히 긴 줄글의 형태로 보여지기에, 끝까지 읽지 않는 이상 핵심이 되는 정보 혹은 사용자가 원하는 정보를 정확하게 파악하기 힘들기 때문이다. 이러한 문제를 해결하기 위해, **본 프로젝트에서는 다양한 제품에 대해 작성된 텍스트 리뷰 데이터들로부터 제품의 잠재적인 제품 속성을 추출하고, 이를 바탕으로 제품 리뷰를 읽는 사용자에게 더 나은 사용자 경험을 제공하고자 한다.** 구체적으로는, 서로 다른 제품들의 리뷰가 혼합된 텍스트 데이터셋을 활용하여 각 제품별 클러스터 및 제품 속성들을 추출하고, 이를 바탕으로 리뷰 토픽(일종의 분류 태그)를 생성하여 그에 따라 리뷰들을 분류해볼 예정이다. 실제 구매 의사 결정 상황에서, 사용자로 하여금 해당 제품의 리뷰를 자신의 구매의사결정 기준에 따라 분류하고, 필요한 정보만 선별적으로 습득할 수 있게 한다. 다시 말해, 제품 리뷰를 자신의 구매 의사결정에 사용자가 보다 효율적으로 활용할 수 있게 하는 것이 궁극적인 목표이다.

2. 왜 이 문제가 흥미로운지에 대한 의견 제시

온라인 리뷰는 상품에 관심이 있는 잠재 고객들에게 필요한 정보를 제공해줌으로써, 그들이 구매결정을 내리는 데 중요한 요소 중 하나로 작용한다. 그러므로 기업 입장에서는 이러한 온라인 리뷰 데이터로부터 유의미한 정보를 추출하여 이를 제품 전략에 반영하는 것이 중요할 것이다. **그러나 제품군마다 유의미한 제품의 속성이 상이하고, '유의미함' 또는 '유의미한 리뷰'의 기준 또한 리뷰를 읽는 잠재 고객마다 다를 수 밖에 없기 때문에, 그 속성 정보를 객관적으로 일일이 추출하기란 결코 쉽지 않은 문제이다. 설령 이를 추출하였다고 하더라도, 각 리뷰를 해당 제품 속성 또는 태그에 맞게 분류하는 것이 쉽지 않다.** 일례로 뷰티앱 '화해'의 경우에는, 앱 내 리뷰 데이터를 활용하기 위하여 기업 차원에서 직접 임의로 약 80개의 화장품 속성을 설정하고 그에 따라 앱 내 모든 화장품 리뷰들에 대해 사람이 일일이 수작업 태깅 작업을 수행하였다. '아마존', '쿠팡' 등과 같이 매우 많고 다양한 제품군을 다루는 온라인 유통업체들은 이러한 태깅의 어려움으로 인해, 단순히 텍스트 리뷰 자체를 리스트업하는 형태로 제품 리뷰를 제공하는 경우가 많다. **본 문제는 이처럼 실제 기업의 문제상황과 데이터셋을 바탕으로 한 실용적인 문제라는 점에서 특히 흥미롭다고 생각되었다.** 더 나아가 이 문제는 비단 리뷰 데이터뿐만 아니라 각 기업 단위에서 발생되는 수많은 텍스트

데이터들을, 기업이 그 필요에 맞게 특정 도메인 또는 토픽을 중심으로)해당 정보만을 효과적으로 추출 및 활용할 수 있게 하는 초석이 될 것으로 생각되기에 더욱 흥미롭게 다가온다. 특히 최근 거대언어모델(LLM)이 큰 주목을 받으며 각 기업에서 이를 잘 효과적으로 활용하기 위한 방안을 도모하고 있는데, 이에 큰 기여를 할 수 있을 것으로 기대된다.

3. 주어진 문제 해결에 있어서 ML의 적합성

3-1. ML 방법론에 대한 서술

제시한 문제를 풀기 위하여 계획한 파이프라인은 다음과 같다.

1) 문서 임베딩

2) 차원 축소

3) 클러스터링

4) 클러스터별 토픽 추출

먼저 문서와 문서 간의 유사도를 비교하기 위해 문서를 임베딩한다. 문서 임베딩 방법에는 단어 빈도수 기반 임베딩, 딥러닝 기반 임베딩 등을 활용할 계획이다. 이후 SVD, PCA 등의 차원 축소 기법을 활용해 임베딩한 문서 데이터의 차원을 축소시키고, 유사한 문서끼리 클러스터링한다. 클러스터링 기법으로는 K-means clustering, agglomerative clustering 등 클러스터링 기법들을 다양하게 활용해보며 적절하다고 판단되는 것을 적용할 계획이다. 클러스터링이 완료된 후에는 클러스터 단위로 토픽 모델링을 적용하여, 유사한 문서들 간에 내재된 토픽을 추출함에 따라 최종적으로 리뷰 태깅에 활용할 수 있도록 할 것이다. 토픽 모델링에는 전통적인 통계 기반 방법론인 LDA(Latent Dirichlet Allocation)를 사용해볼 계획이지만, 사전 조사 결과 LDA도 다소 한계점이 있는 기법이라고 판단되어 추가적으로 관련 논문을 찾아보며 보완할 만한 토픽 모델링 기법들이 있다면 적용해보려고 한다.

3-2. 해당 ML방법론 선택 이유

리뷰 데이터는 다양한 상품 및 주제가 혼재된 텍스트 데이터이므로, 사용자에게 보기 쉽게 제공하기 위해서 의미가 유사한 것끼리 묶어 분류해내는 작업이 필요하다. 그렇지만 이러한 작업은 사람이 일일이 해내거나 분류 기준 자체를 정의하기가 어렵기에, 기계학습 방법론을 적용해볼 수 있을 것이다. 그리고 앞서 언급했듯이, 제품마다 유의미한 속성이 모두 다르고, 이 속성 또한 사람이 정해주어야 한다는 문제점 때문에 레이블을 필요로 하는 지도학습 기반의 분류 문제로 접근하기에는 무리가 있다. 따라서 비지도학습 기반의 클러스터링 기법을 활용, 별도의 레이블 없이도 객관적이고 통일성 있게 글을 분류할 수 있도록 하려 한다. 그 과정에서 문서를 벡터화하기 위한 문서 임베딩 기법과, 클러스터링 시 성능을 높이기 위해 차원 축소 기법을 추가적으로 활용할 예정이다.

리뷰 내용에 따라 제품마다 추출된 제품 속성들이 상이하기 때문에, 리뷰들을 대표하는 속성을 추출하는 것 역시 쉽게 해결할 수 없는 문제이다. 속성을 규칙 기반이나 사람의 판단에 의지해 추출해내거나, 모든 리뷰에 대해 일괄적으로 같은 속성을 사용하는 등 다양한 방법론이 있지만, 토픽

모델링을 이용해 데이터에 맞는 속성을 추출해내어 각 리뷰를 태깅하는 것이 더 효율적이고 유연한 방법이라고 판단하였다. 뿐만 아니라, 토픽 모델링을 적용하여 리뷰들에 잠재된 주제를 추출해낼 수 있다면, 도메인에 구애받지 않고 상품을 대표하는 속성들을 구할 수 있을 것이다.

4. 대략의 프로젝트 수행 일정

번호	작업 제목	시작일	마감일
1	**Topic & Data**		
1.1.1	주제 선정 및 일정 공유	23년 9월 14일	23년 9월 21일
1.1.2.	주제 확립 및 데이터셋, 모델 확보	23년 9월 21일	23년 9월 27일
1.2.1.	데이터 전처리 방법 구상	23년 9월 27일	23년 10월 3일
1.2.2.	데이터 전처리 완료	23년 10월 1일	23년 10월 8일
1.3.	metric 조사 및 세팅	23년 10월 8일	23년 10월 17일
2	**Method & Experiments**		
2.1.	method 구상 완료	23년 9월 26일	23년 10월 3일
2.2.	method 파이프라인 정립 및 세팅 완료	23년 10월 1일	23년 10월 8일
2.3.1.	1차 experiment 진행 및 보고	23년 10월 6일	23년 10월 11일
2.3.2.	2차 experiment 진행 및 보고	23년 10월 12일	23년 10월 17일
2.4.	1, 2차 experiment 결과 확인 및 분석	23년 10월 17일	23년 10월 20일
3	**Report**		
3.1	프로포절 작성	23년 9월 26일	23년 9월 28일
3.2.	related works 정리	23년 9월 28일	23년 10월 6일
3.3.	mid-report 작성	23년 10월 8일	23년 10월 12일
3.4.	reference 정리	23년 10월 17일	23년 10월 19일
3.5.1.	final-report 초안 작성	23년 10월 13일	23년 10월 19일
3.5.2.	final-report 퇴고 후 제출	23년 10월 19일	23년 10월 26일

5. 예상되는 결과

본 프로젝트에서는 클러스터링과 토픽 모델링 두 단계를 거치며, 실험 과정에서 여러가지 ablation을 시도할 예정이다. 토픽 모델링 기법으로 LDA를 활용할 것이며, 파이프라인의 각 phase에서 여러 모델들을 적용해보며 각 방법론 간의 비교 분석을 진행할 계획이다. 이 때, 레이블이 붙어 있는 테스트 데이터셋을 마련하여, 실제 토픽들이 consistent하게 나타나는지, 잘 클러스터링이 되는지, 유사한 토픽 간의 similarity를 계산하는 등을 통해서 성능을 비교할 것이다. 그 과정에서 토픽 클라우드나 히트맵을 통해 시각화하여 더 분명하게 나타낼 수 있다.

그 결과 리뷰가 입력으로 들어오면 주요 단어와 주요 토픽 등을 확인할 수 있고, 이를 기반으로 데이터 구조 및 토픽 간의 상호 관계도 이해할 수 있다. 예를 들어, 문제 정의에서 언급하였던 '아마존'이나 '쿠팡'에서 판매하는 의류 상품 리뷰 데이터가 입력으로 들어올 경우, 디자인과 가격, 색상 등 옷 데이터에 맞는 속성을 추출해낼 수 있다. 혹은, 호텔 리뷰 데이터를 입력한다면, 전망, 가격, 서비스 등을 출력하도록 할 수 있다. 이런 식으로 실제 기업의 리뷰 혼합 데이터셋을 적용해본다면, 본 프로젝트에서 문제로 지적한 점들을 해결하리라 기대하고, 특히, 특정 주제의 문서를 분류하는 데에서 정확성과 효율성을 보다 더 향상시킬 수 있을 것으로 가설을 세울 수 있다.