온라인 리뷰 특화 한국어 자연어 처리 모델 구현

: ReBERT, ReELECTRA

1 서론

인터넷과 스마트폰의 발전으로 인터넷 쇼핑에 대한 접근성이 늘어나면서 원하는 상품을 언제 어디서나 구매할 수 있는 환경이 되었다. 인터넷 쇼핑몰을 통해 상품을 구매하는 것은 오프라인에서 구매하는 것 보다 시간과 장소에 영향을 받지 않고 편리하게 이용할 수 있다는 장점이 있다. 위와 같은 오프라인에서 온라인으로서의 소비 채널의 변화는 전자상거래 시장의 성장을 이끌었다. 통계청에서 발표한 자료에 따르면 국내 전자상거래 시장 규모는 2018 년부터 최근 5 년간 연평균 성장률이 15.1%로 글로벌 e 커머스 시장의 연평균 성장률인 14.6%를 넘어선다. 한국 e 커머스 시장 규모는 상기한 바와 같은 성장률로 2018 년에는 113 조에 불과했던 시장이, 2023 년 228 조에 달한다고 한다. 시장 규모가 커짐에 따라 신규 집입자, 대체품의 위협은 커지고 산업 내 경쟁자 간 경쟁의 강도는 높아진다. 이에 대응하기 위해 기업은 단순 가격 경쟁이 아닌 차별화 전략을 통해 경쟁력을 유지해야 한다(Porter, M. E. “Five Forces Framework). 현대 경영학의 대가인 피터 드러커(Peter F.Drucker)는 "기업의 목적은 고객을 창조하는 것이다”라고 마케팅을 정의하였다. 마케팅은 단순 물건을 파는 것을 넘어, 소비자의 충족되지 못한 욕구를 발견하고 이를 해결하는 방법을 제시하여 판매를 자연스럽게 성사시키는 것이 마케팅이라는 것이다. 피터 드러커가 고객의 중요성을 강조한 이래로 현재까지 많은 기업들은 고객만족경영, 소비자중심경영을 기업의 중심가치로 내세우며, 고객만족을 증대시키기 위한 다양한 전략방안들을 내오고 있다[1].

온라인 고객 리뷰(이하 OCR)는 온라인 쇼핑몰을 통해 직접 제품을 구매한 고객이 남긴 것으로 특히, 온라인쇼핑에서는 상품 및 구매자를 직접 보지 못하는 특성 때문에 대다수의 소비자들은 앞선 구매자들의 상품평이나 구매후기에 의존해 구매여부를 판단하게 된다. 즉 다른 소비자가 제공하는 제품 사용경험이나 사용후기 등과 같은 정보를 보다 더 적극적으로 탐색하여 구매판단에 반영하며(Peterson, etal., 2003), 소비자들은 온라인 상에서의 제품에 대한 리뷰를 합리적인 구매결정을 내리기 위한 중요한 정보로 활용하고 있다. 실제, 많은 연구들은 리뷰의 수와 리뷰의 내용이 제품 판매에 중요한 영향을 미친다는 점을 실증적으로 보여주고 있다[2]. 이렇듯 OCR 은 직접 제품을 구매, 경험한 고객 만족의 결과물이자 온라인 쇼핑몰 잠재 고객의 구매 의사 결정에 지대한 영향을 준다[3]. 특히, 부정리뷰는 잠재 고객들에게 제품과 기업에 대한 부정적인 초두 효과를 초래한다. 하지만, 부정리뷰는 브랜드가 상품을 개선할 수 있는 방향을 제공하는 지표가 되기도 한다.

온라인 고객 리뷰 분석을 통해 기업은 제품의 개선, 마케팅 전략의 최적화, 경쟁사 분석 또한 가능하여 많은 기업과 제품들이 난립하는 오늘날의 온라인 시장에서의 높은 고객생애가치를 위해, 타 기업과의 차별점이 필요한 기업에서 온라인 고객 리뷰 분석은 기업 전략 개선의 원천이고 그 분석은 온라인 시장에서의 경쟁 우위를 확보하기 위한 필수적 과제라 할 수 있다.

이와 같은 필요성은 단순히 OCR 을 수집하는 수준을 넘어 체계적 분석, 즉 고객 만족의 결과물인 리뷰를 정량적, 정성적으로 해석하여 기업의 차별화 전략과 고객만족경영을 강화하여야 한다. 자연어 처리 모델(NLP)을 이용한 분석이 방대한 기업의 리뷰 데이터를 체계적으로 해석해 차별화 전략과 고객만족경영을 강화하는 핵심 도구로 사용될 수 있다. 현재 한국어 NLP 분야에서는 KcBERT, KcELECTRA 등 범용 모델이 존재하나, 이들 모델은 일반적인 한국어 말뭉치에 기반해 학습된 범용 모델로서 '리뷰'라는 특수한 도메인에 최적화되어 있지 않다.

따라서 본 프로젝트에서는 온라인 고객 리뷰에 특화된 한국어 NLP 모델을 새롭게 구현하였다. 본 모델의 긍정, 부정 분류 결과는 상품의 품질 개선, 물류 서비스의 개선, 브랜드 이미지 관리 등 기업의 핵심 경영 활동에 직접 활용될 수 있으며, 동시에 쿠폰, 적립금, 사은품 제공과 같은 판촉 활동과 연계되어 리뷰 활성화를 촉진할 수도 있다. 이러한 일련의 과정은 제품을 구매한 고객의 리뷰가 잠재 고객을 실제 고객으로 끌어들이고, 그로 인해 활성화된 리뷰가 다시 더 많은 긍정 리뷰를 생성하는 선순환 구조를 형성한다. 이러한 선순환을 실현하기 위해 본 프로젝트에서는 온라인 고객 리뷰 분석에 특화된 한국어 NLP 모델을 제안하며, 해당 모델의 구조와 학습 과정, 그리고 기존 범용 모델 대비 갖는 차별성을 설명하고자 한다.

2 관련 연구

2.1 감성 분석(Sentiment Analysis)

감성 분석(Sentiment Analysis)은 자연어 처리(Natural Language Processing, NLP) 분야에서 널리 활용되는 분석 기법으로, 텍스트 데이터를 기반으로 메시지의 감정적 어조를 파악하는 방법이다. AWS 에서는 이를 "디지털 텍스트를 분석하여 메시지의 감정적 어조가 긍정적인지, 부정적인지 또는 중립적인지를 확인하는 프로세스"라고 정의한다[4]. 즉, 감성 분석은 언어 속 감정 신호를 정량화 하여 데이터 기반 의사 결정을 지원하는 핵심 기술로 간주된다.

감성 분석의 접근 방법은 크게 사전 기반 접근 방식과 머신러닝 기반 접근 방식으로 구분된다. 사전 기반 접근은 긍정·부정을 나타내는 단어 리스트(감성 사전, sentiment lexicon)를 이용해 문서의 감정을 계산하는 방식으로, 구현이 간단하고 직관적이다. 그러나 문맥, 반어법, 다의어 등 복잡한 언어적 특성을 충분히 반영하지 못한다는 한계가 있다[5], [6]. 반면 머신러닝 기반 접근은 감성 레이블이 부여된 데이터셋을 이용해 모델이 직접 패턴을 학습하도록 하며, Naive Bayes, SVM(Support Vector Machine), 그리고 딥러닝 모델(CNN, GRU, BERT, ELECTRA) 등이 대표적으로 사용된다. 이러한 접근은 대량의 데이터 학습을 통해 문맥적 의미를 이해하고, 보다 정교한 감성 분류를 수행할 수 있다[7].

최근에는 인공지능 기술의 발전과 함께 대규모 언어모델(LLM, Large Language Model)을 활용한 감성 분석 연구도 활발히 진행되고 있다. LLM 은 방대한 양의 데이 터를 사전학습하여 문맥과 감정의 뉘앙스를 정밀하게 해석할 수 있으며, 특정 도메인에 맞게 커스터마이징이 가능하다고 보고되었다[8].

2.2 BERT

BERT(Bidirectional Encoder Representations from Transformers)는 구글이 제안한 사전학습(Pre-training) 언어 모델로, 문맥을 양방향으로 이해하는 특징을 가진다. 기존의 RNN 이나 LSTM 계열 모델은 문장을 한쪽 방향(좌→우 또는 우→좌)으로만 순차적으로 처리하기 때문에 문맥의 일부 정보만 반영할 수 있었으나, BERT 는 Transformer 의 Encoder 구조를 기반으로 문장 전체를 동시에 분석하여 양방향 문맥을 모두 고려할 수 있다. 또한 BERT 는 모델 크기에 따라 base 와 large 두 가지 주요 모델을 제공한다. BERT-base 는 12 개의 Transformer 레이어, 12 개의 어텐션 헤드, 768 차원의 히든 크기를 가지며, 상대적으로 가벼운 구조로 다양한 응용에 활용된다. 반면, BERT-large 는 24 개의 레이어, 16 개의 어텐션 헤드, 1024 차원의 히든 크기를 갖추어 더 강력한 표현력을 제공하지만, 연산 비용이 크다는 특징이 있다.

BERT 의 핵심 개념은 양방향 문맥 이해(Bidirectional Context Understanding)와 사전학습 후 미세조정(Pre-training & Fine-tuning)의 두 단계 학습 과정에 있다. 먼저 사전학습 단계에서는 두 가지 주요 목표인 Masked Language Model(MLM)과 Next Sentence Prediction(NSP)을 통해 언어의 의미적 표현을 학습한다. MLM 은 입력 문장 내 일부 토큰을 [MASK]로 치환한 후, 해당 위치의 원래 단어를 예측하도록 학습하는 방식이다. 전체 입력 토큰 중 약 15%를 무작위로 선택하여 마스크 처리하며, 이 중 80%는 [MASK], 10%는 임의의 다른 단어, 나머지 10%는 원래 단어를 그대로 유지한다. 이러한 방식은 모델이 문장의 양방향 문맥 정보를 모두 활용하여 단어의 의미를 추론하도록 유도한다. NSP 는 두 문장이 실제로 연속된 문장인지 여부를 예측하는 문장 간 관계 학습(Task)이다. 학습 데이터의 절반은 실제로 이어지는 문장 쌍(positive pair), 나머지 절반은 무작위로 추출된 비연속 문장 쌍(negative pair)으로 구성된다. 이를 통해 모델은 문장 간 논리적 연결성과 담화적 맥락을 학습하게 된다.

다만, 우리가 만들 리뷰 긍부정 예측 모델은 NSP 가 DAPT 학습에 불필요하고, 논문상 NSP 학습을 하지 않는 BERT 모델도 성능이 MLM 만 사용한 모델과 성능차이가 나지 않고, 오히려 더 떨어진다는 논문을 기반으로 해서, 이번 Re-BERT 모델에서는 사용하지 않는다.

2.3 ELECTRA

ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)는 구글이 제안한 사전학습(Pre-training) 언어 모델로, 기존 BERT 의 MLM(Masked Language Modeling) 방식을 대체하기 위해 개발되었다. ELECTRA 의 핵심 아이디어는 RTD(Replaced Token Detection) 방식을 통해 모델이 더 효율적으로 언어적 표현을 학습하도록 하는 것이다.

ELECTRA 의 학습 구조는 Generator(생성기)와 Discriminator(판별기)로 구성된다.

Generator 는 BERT 와 동일한 구조를 가지며, 입력 문장에서 일부 토큰(약 15%)을 마스킹한 뒤 이를 예측한다. 예측 결과로 생성된 토큰들은 원래의 문장 내 마스킹된 위치에 다시 삽입되어, 부분적으로 '가짜 토큰(Replaced Token)'이 포함된 새로운 문장을 만든다.

Discriminator 는 이 수정된 문장을 입력으로 받아, 각 토큰이 원래의 진짜 토큰인지 혹은 Generator 가 생성한 가짜 토큰인지를 판별하는 이진 분류 과제를 수행한다. 즉, ELECTRA 는 BERT 처럼 일부 마스킹된 토큰만을 예측하는 대신,

문장 내 모든 토큰에 대해 학습 신호를 부여함으로써 훨씬 더 효율적인 학습이 가능하다.

이 구조 덕분에 ELECTRA 는 동일한 학습 데이터와 계산량으로도 BERT 보다 빠르고 높은 성능을 달성할 수 있다. 실제로 Generator 는 상대적으로 작은 모델로 구성되어 계산 비용을 절감하며, Discriminator 가 문맥적 일관성을 정밀하게 파악하도록 유도한다. 이러한 효율성은 ELECTRA 가 제한된 자원 환경에서도 뛰어난 성능을 보이는 이유이기도 하다.

2.4 한국어 자연어 처리 모델

현재 공개된 한국어 자연어 처리 모델로는 KoBERT, KoELECTRA, KcBERT, KcELECTRA 등 이 있다.

KcBERT/KcELECTRA는 뉴스 위키뉴스, 위키백과, 국립 국어원에서 제공한 모두의 말뭉치 등 문어체 데이터셋을 기반으로 학습한 모델이다.

KcBERT / KcELECTRA는 일상 대화, 신조어 등이 주로 사용되는 온라인 리뷰에서 수집한 구어체 데이터셋을 기반으로 학습한 모델이다.

본 프로젝트는 온라인 쇼핑몰 리뷰 데이터를 대상으로 감성(긍정 / 부정) 분류를 수행하는 것을 목표로 한다. 이를 위해 구어체 분석에 우수한 성능을 보이는 KcBERT / KcELECTRA의 특성을 참고하여 학습을 진행한다.

2.5 DAPT : 도메인 적응 사전 학습

DAPT 란 도메인 적응 사전학습(Domain-Adaptive Pre-training)이다. DAPT 는 특정 도메인의 데이터를 활용하여 기존 사전학습된 언어 모델의 성능을 개선하는 기법으로[9], 기존의 사전 훈련된 언어 모델이 가지고 있는 일반적인 언어 이해 능력을 특정 도메인에 맞게 최적화하는 데 중점을 둔다[10]. 이러한 최적화 과정은 본 프로젝트 모델이 리뷰라는 특정 도메인에서 미묘한 언어적 차이와 감정적 차이를 이해하고, 기존 한국어 자연어 처리들 보다 분류에 있어 정확한 성능을 발휘하게 한다.

이를 위해 BERT, ELECTRA 자연어 처리 모델에 한국어를 사전학습 시킨 후 DAPT 를 이용하여 리뷰 특화 한국어 자연어 처리 모델을 구현한다.

3 데이터셋 구성 및 정제

3.1 데이터 세트 구성

본 프로젝트에서는 사전 학습(pre-training) 데이터, DAPT 데이터, 파인 튜닝(fine-tuning) 데이터, 그리고 분류를 위한 크롤링 데이터로 구성한다.

사전 학습 데이터는 모델의 일반적 언어 이해 능력을 확보하기 위해 다양한 도메인에서 수집하였다. 뉴스 리뷰 150만 건(GitHub), 일상 구어체 20만

건(AIHub), 뉴스 기사 80만 건(AIHub), 한국문화 기사 10만 건(AIHub), 웹사이트 설명 10만 건(AIHub), 게임 리뷰 10만 건(GitHub), 영화 리뷰 20만 건(GitHub)을 포함한다. 이를 통해 총 약 300만 건의 텍스트 데이터를 사전학습에 사용한다.

DAPT 데이터는 강화시킬 특정 도메인의 데이터를 활용하여 사전 훈련된 언어 모델의 성능을 개선 해야 한다. 본 프로젝트의 목적은 리뷰 특화 모델을 만드는 것 이기에 쇼핑몰 리뷰 22만건(AIHub)을 사용한다.

파인 튜닝 데이터는 온라인 쇼핑 리뷰에서의 모델 성능을 향상시키기 위해 네이버 쇼핑 리뷰 20만 건을 사용한다.

크롤링 데이터는 온라인 쇼핑 플랫폼(무신사)에서 수집한 최신 리뷰 데이터로, 수집 항목은 사용자 ID, 리뷰 텍스트, 별점, 상품명, 가격, 리뷰 수, 조회 수, 누적 판매 수, 할인율 등이며, 카테고리별 후기 많은 순으로 정렬한 뒤 최신순으로 수집하였다. 구체적으로 아우터(패딩, 블레이저), 상의(맨투맨, 후드티, 반소매 티셔츠), 하의(트레이닝 팬츠, 슬랙스, 데님)로 구성되어 있으며, 각 상품당 5천건, 각 분류 별 3만 건씩 총 12만 건을 수집하였다.

이와 같은 데이터셋 구성은 모델이 일반적 언어 이해 능력과 도메인 특화 성능을 동시에 확보할 수 있도록 설계, 수집되었다.

3.2 데이터 정제

본 프로젝트에서 활용한 데이터는 상기한 바와 같이 뉴스 기사, 리뷰, 구어체 발화 등 다양한 출처와 도메인에서 수집되었기 때문에 원천 데이터의 형식과 품질이 상이하다. 이러한 이질성은 토큰화 오류, 희귀 토큰 폭주, 의미적 중복, 학습 안정성 저하로 이어져 일반화 성능을 손상시킬 수 있다

따라서 수집 데이터 전반에 일관된 정제·통일 규칙을 적용하여 수집 데이터를 재구성하는 과정이 필수적이다. 기존 한국어 구어체 자연어 처리 모델인 KcBERT의 정제 규칙을 참고하였다.

첫째, 한글(자모 포함)과 영어, 특수문자, 유니코드 이모지, 숫자를 제외한 모든 비표준 문자는 제거한다. 이를 통해 모델 입력에서 불필요한 잡음을 최소화하고 토큰화 과정의 안정성을 확보한다.

둘째, 온라인 리뷰와 댓글에서 빈번히 나타나는 중복 문자열을 정규화하였다. 예를 들어 ‘ㅋㅋㅋㅋ’와 같은 반복 표현은 soynlp 라이브러리를 활용하여 두 글자 ‘ㅋㅋ’로 축약한다. 이는 의미적 신호는 유지하면서 데이터의 불필요한 변동성을 줄이는 효과를 가진다.

셋째, 글자 단위 10자 이하의 초단문은 제거한다. 지나치게 짧은 문장은 정보량이 부족하여 학습에 기여하지 못할 가능성이 크므로, 이를 배제함으로써 모델이 충분한 문맥을 학습할 수 있도록 한다.

넷째, 동일하거나 유사한 문장이 과도하게 포함되는 것을 방지하기 위해 중복 문장을 제거한다. 이는 데이터셋의 다양성을 확보하고 특정 문장에 대한 과적합을 예방하는 데 기여한다.

3.3 단어사전

본 프로젝트에서는 WordPiece 알고리즘을 사용하여 단어 사전(vocab)을 생성하였다. WordPiece 는 Google 에서 BERT 를 사전학습하기위해 개발한 토큰화 알고리즘으로, Hugging Face 에서 제공하는 Tokenizers 라이브러리를 통해 사용할 수 있다. 위 알고리즘은 문장을 최소 의미 단위(서브워드)로 나누는 규칙을 학습하는 알고리즘으로 함께 자주 등장하는 문자를 합쳐 서브워드 단위로 단어 사전을 생성한다.

출현 빈도가 높은 문자 조합을 병합하여 서브워드 단위로 단어 사전을 생성한다. Tokenizer 는 입력받은 문장을 모델이 처리할 수 있도록 학습된 단어 사전을 기반으로 문장을 토큰화하고, 해당 토큰과 맵핑된 정수 인덱스로 변환한다.

vocab parameter

- vocab_size : 30000 -> 일반적으로 30000~32000 정도로 설정한다.
- limit_alphabet : 3000 -> 한국어 개별 문자, 특수 문자 등의 최대 개수를 제한하여 Out of Vocab 발생을 최소화한다.

4 프로젝트 설계

4.1 ReBERT

본 프로젝트에서는 제한된 환경상 GPU(Graphic Processing Unit)의 부담을 줄이기 위해 KcBERT-base 모델 대비 약 절반 규모의 설정(레이어 수 : 6 개, 헤드 수 : 8 개, 히든 : 512)을 적용하여 학습하였다. 사전학습(Pre-training) 단계에서는 일반적인 BERT 모델 튜닝에서 사용하는 값으로 적용하였으며, 학습 주기(epoch)는 언어적 표현 학습에 충분한 반복 횟수로 판단되는 3 epoch 을 설정하였다. 또한, 사전학습 단계에서는 대규모 데이터셋을 빠르게 학습하기 위해 상대적으로 높은 학습률을 적용하였다. 반면, 도메인 적응 사전학습(DAPT) 단계에서는 이미 학습된 모델의 가중치를 유지하면서 도메인 데이터에 적응하도록 해야 하므로, 과적합을 방지하고 안정적인 수렴을 유도하기 위해 보다 낮은 학습률을 적용하였다.

4.1.1 사전학습

Epoch : 3, Batch Size : 32, Learning Rate : 1e-4

4.1.2 DAPT

Epoch : 3, Batch Size : 32, Learning Rate : 5e-5

4.1.3 파인튜닝

Epoch : 3, Batch Size : 32, Learning Rate : 5e-5

4.2 ReELECTRA

본 프로젝트에서는 상기한 이유로 Google 의 ELECTRA-small 과 동일한 설정(config)으로 모델을 학습하였다. Generator 와 Discriminator 의 학습률(learning

rate)은 ELECTRA 의 구조적 특성을 고려하여 차별적으로 설정하였다. 특히, Generator의 학습률을 Discriminator의 절반 수준으로 낮게 조정한 이유는, Generator가 마스킹된 토큰을 지나치게 정확하게 예측할 경우 Discriminator 가 해당 토큰의 진위 여부를 판별하기 어려워져 판별 학습의 효과가 저하되기 때문이다.

4.2.1 사전학습

Epoch : 3, Batch Size : 16, Learning Rate : 1e-4

4.2.2 DAPT

Epoch : 3, Batch Size : 16, Learning Rate : 5e-5

4.2.3 파인튜닝

Epoch : 3, Batch Size : 16, Learning Rate : 5e-5

4.3 모델 평가 방법

본 실험은 총 네 가지 모델(KcBERT, ReBERT, KcELECTRA, ReELECTRA)의 정확도(Accuracy) 와 F1 점수(F1 Score)를 비교하여 모델 성능을 평가하는 방식으로 진행한다. 파인튜닝 데이터는 긍부정의 비율이 약 6:4 로 다소 불균형한 분포를 보였으나, 학습 과정에 큰 영향을 미치지 않아, 별도의 비율 조정은 수행하지 않는다.

1. Accuracy : 전체 예측 중 정확한 예측의 비율
2. F1 Score : 정밀도와 재현율의 조화 평균

4.4 결론

4.4.1 기존 모델 대비 차별점

본 연구는 온라인 쇼핑몰 리뷰 데이터를 대상으로 긍정·부정 감성 분류를 수행하기 위해, 리뷰 도메인에 특화된 한국어 자연어처리 모델 ReBERT 와 ReELECTRA 를 제안한다. 서론에서 언급한 바와 같이, 온라인 고객 리뷰는 소비자의 구매 의사 결정에 중요한 영향을 미치며, 기업의 제품 개선 및 마케팅 전략 수립에 활용 가능한 핵심 데이터이다. 그러나 기존 범용 한국어 모델(KcBERT, KcELECTRA 등)은 일반 문체 중심으로 학습되어 리뷰 특유의 구어체적 표현과 감정적 뉘앙스를 충분히 반영하기 어렵다는 한계가 있다.

프로젝트 모델은 기존 한국어 자연어 처리 모델 대비 적은 학습 데이터와 경량화 된 구조를 사용하지만 감성 분류 성능에서는 유사하거나 높은 수준의 결과를 목표로 한다..

이는 연산 비용 절감(Computational Cossst Reduction)과 도메인 적합성 확보(Domain Adaptation) 측면에서 충분한 경쟁력을 가진다고 평가할 수 있다.