

Final Project Deep Learning

Min su Jo

ID : 2021122062

제품 리뷰 기반 평점 예측 모델 보고서

1. 서론

평소 나는 e커머스를 자주 이용하면서 제품을 구매할 때 리뷰를 꼼꼼히 살펴보는 편이다. 그런데 막 출시된 상품의 경우에는 리뷰 수가 적어, 신뢰할 수 있는 판단 기준이 부족하다는 생각이 들었다. “초기 리뷰 몇 개만으로도 이 제품의 미래 평점이 예측 가능하지 않을까?”라는 의문에서 본 프로젝트는 출발했다.

본 프로젝트에서는 Amazon의 리뷰 데이터를 활용하여, 제품의 초기 리뷰를 기반으로 향후 평균 평점을 예측하는 딥러닝 모델을 구축하고자 한다. 이를 통해 리뷰 텍스트에 담긴 정보가 장기적인 제품 평가를 예측하는 데 얼마나 효과적인지를 분석하고, 특히 초기 5~10개의 리뷰만으로도 어느 정도 예측력을 확보할 수 있는지 확인하고자 하였다.

이 연구는 사용자 정보나 상품의 메타 데이터없이, 자연어 처리 기법과 딥러닝 모델만을 활용하여 평균 평점을 예측하는 것을 핵심 목표로 한다. 이는 이커머스 플랫폼에서 신상품 출시 초기에 시장 반응을 빠르게 파악하는 데 실질적인 도움을 줄 수 있을 것으로 기대된다.

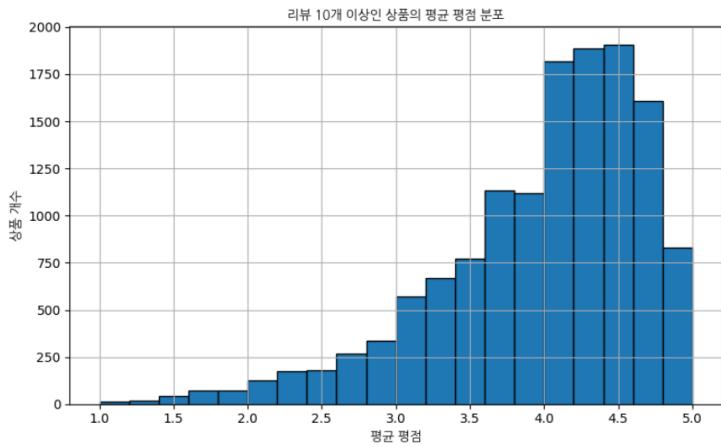
본 연구는 다음의 세 가지 질문에 답하고자 한다:

1. 초기 리뷰에 담긴 감성적·언어적 특징이 이후 평점 변화와 얼마나 상관관계가 있는가?
2. 소수의 텍스트 데이터만으로도 기본적인 예측 정확도(MAE 0.5 이내)를 달성할 수 있는가?
3. 간단한 구조의 모델(예: 임베딩 + LSTM)만으로도 충분한 학습이 가능한가?

이처럼 본 프로젝트는 현실적인 조건 내에서 텍스트 데이터의 예측 잠재력을 검증하고, 그 결과를 통해 실제 모델 적용 가능성은 판단하고자 한다.

2. 전처리

본 연구에서는 2023년 기준으로 수집된 Amazon 제품 리뷰 데이터셋을 활용하였으며, 모델 학습을 위한 데이터 전처리 과정을 다음과 같이 수행하였다. 해당 데이터는 asin (제품 고유 식별자), text (리뷰 텍스트), rating (평점), time_stamp (리뷰 작성 시점) 등의 항목을 포함하고 있으며, 본 프로젝트에서는 asin과 text를 중심으로 분석을 진행하였다.



제품별 평균 평점 분포를 살펴보면, 4.0~5.0점 구간에 가장 많은 상품이 분포하고 있으며, 3.5~4.0점과 4.5~5.0점 구간도 상당한 비중을 차지하고 있다. 이는 대부분의 제품이 높은 평점을 받고 있음을 시사한다.

우선 전체 데이터 중 제품별 리뷰 수가 10건 이상인 경우만 선별하여 최소한의 학습 샘플 수를 확보하였다.

- 2.1 데이터 선별: 전체 데이터 중 제품별 리뷰 수가 10건 이상인 경우만 선별하여 최소한의 학습 샘플 수를 확보하였다.
- 2.2 초기 리뷰 추출: 각 제품별 최초 리뷰를 기준으로 6개월 이내에 작성된 리뷰들만을 초기 리뷰로 간주하였다. 이는 출시 직후의 소비자 반응만을 반영하고자 한 것으로, 장기적인 평점 변화에 영향을 줄 수 있는 후기 리뷰는 제외하였다.
- 2.3 데이터셋 구성: 필터링된 각 제품의 초기 리뷰들을 입력으로 설정하고, 해당 제품의 전체 평균 평점을 예측 대상으로 구성하였다.
- 2.4 텍스트 전처리: 리뷰 본문은 모두 소문자로 변환하였으며, URL, 특수문자, 불용어(stopwords)를 제거하여 모델 학습에 불필요한 요소를 배제하였다. 이후 Keras의 Tokenizer를 사용해 단어 인덱스를 구성하였고, 등장 빈도 기준 상위 10,000개의 단어만을 사용하였다. 이렇게 생성된 정수 시퀀스는 pad_sequences 함수를 통해 고정된 길이(예: 500)로 패딩 처리하여 모델에 입력할 수 있는 형태로 변환하였다.
- 2.5 훈련/검증 데이터 분할: 입력 데이터(X_{train}, X_{val})와 타깃값(y_{train}, y_{val})을 구성하는 과정에서는 제품 단위로 시계열 순서에 따라 리뷰 시퀀스를 정렬한 후, 이를 80:20의 비율로 훈련 및 검증 데이터셋으로 분할하였다.

이와 같은 전처리 과정을 통해, 실제 제품 출시 초기의 제한된 리뷰 정보만을 기반으로 후속 평점을 예측하는 문제로 구조화된 입력 데이터를 구축할 수 있었으며, 이 데이터는 LSTM 기반 딥러닝 모델의 학습에 활용되었다.

3. 모델링

Model: "functional_1"			
Layer (type)	Output Shape	Param #	Connected to
input_text (Inputlayer)	(None, 500)	0	-
embedding (Embedding)	(None, 500, 64)	640,000	input_text[0][0]
lstm_2 (LSTM)	(None, 500, 64)	33,024	embedding[0][0]
dropout_1 (Dropout)	(None, 500, 64)	0	lstm_2[0][0]
lstm_3 (LSTM)	(None, 500, 32)	12,416	dropout_1[0][0]
attention (Attention)	(None, 500, 32)	0	lstm_3[0][0], lstm_3[0][0]
context_vector (Lambda)	(None, 32)	0	attention[0][0]
output_rating (Dense)	(None, 1)	33	context_vector[0][0]
Total params:		685,473	(2.61 MB)
Trainable params:		685,473	(2.61 MB)
Non-trainable params:		0	(0.00 B)

본 연구에서는 리뷰 텍스트를 기반으로 미래 평점을 예측하기 위해, LSTM(Long Short-Term Memory)과 Attention 메커니즘을 결합한 딥러닝 모델을 설계하였다. 전체 모델은 입력 임베딩부터 시계열 처리, 주의(attention) 기반 문맥 벡터 추출 까지의 구조를 포함하고 있으며, 모델의 총 파라미터 수는 약 68만 개(685,473개)이다.

3.1 입력 및 임베딩 단계

모델의 입력은 최대 길이 500의 정수 시퀀스 형태의 리뷰 데이터이다. 각 정수는 사전 구축된 단어 인덱스를 의미하며, 해당 단어는 임베딩 레이어를 통해 **64차원 임베딩 벡터**로 변환된다. 이 임베딩 계층은 단어 간 의미적 유사성을 반영하는 벡터 공간을 학습하며, 전체 파라미터 수는 약 64만 개(10,000개의 단어 × 64차원)이다.

3.2 LSTM 기반 시계열 처리

임베딩 벡터 시퀀스는 두 개의 LSTM 계층을 거쳐 시계열적으로 처리된다. 첫 번째 LSTM은 각 시점에 대해 **64차원의 출력 시퀀스**를 생성하며, `return_sequences=True` 설정을 해 전체 시퀀스를 다음 계층으로 전달한다. 이어지는 두 번째 LSTM은 이 시퀀스를 받아 **32차원의 출력 시퀀스**로 압축한다. 각 LSTM 계층 사이에는 **Dropout** 레이어를 삽입하여, 학습 시 과적합을 방지하고 모델의 일반화 성능을 확보하였다.

3.3 Attention 레이어 및 문맥 벡터 추출

마지막 LSTM 계층의 출력에 대해 **Attention 레이어**를 적용하여, 시퀀스 내에서 중요한 단어에 더 큰 가중치를 부여하였다. 이 메커니즘은 입력 시퀀스 전체에서 중요한 단어에 집중함으로써 모델이 보다 효과적으로 의미를 추출할 수 있게 한다. Attention 가중치는 학습을 통해 자동으로 조정되며, 이는 단순 평균보다는 더 정교한 문맥 벡터(Context Vector)를 생성하는 데 기여한다.

이러한 Attention 구조는 특히 긴 시퀀스에서 중요한 정보가 희석되지 않도록 하며, 예측 성능을 향상시키는 데 실질적인 역할을 한다. 본 모델에서는 `tf.keras.layers.Attention` 모듈을 사용하여 Query와 Key를 동일한 시퀀스로 설정하는 **Self-Attention** 구조를 구현했다.

4. 결과 및 해석

본 섹션에서는 구축된 딥러닝 모델의 학습 결과와 예측 성능을 분석하고, 주요 예측 사례를 통해 모델의 작동 방식을 해석한다.

4.1. 모델 학습 과정 분석

모델 학습은 총 10 Epoch에 걸쳐 진행되었으며, 손실(Loss)과 평균 절대 오차(MAE)의 변화를 통해 학습 과정을 모니터링 했다.

학습 추이: 학습 초반 (Epoch 1-4)에는 손실(loss)과 평균 절대 오차(mae)의 감소가 완만하게 진행되었다. 그러나 Epoch 5부터는 검증 손실(val_loss)과 검증 평균 절대 오차(val_mae)가 현저히 감소하며 모델이 데이터의 패턴을 빠르게 파악하고 성능 개선을 이루었음을 확인할 수 있다.

4.2. 예측 성능 지표

모델의 최종 예측 성능은 검증 데이터셋에 대한 평균 절대 오차(MAE)와 평균 절대 백분율 오차(MAPE)를 통해 평가했다.

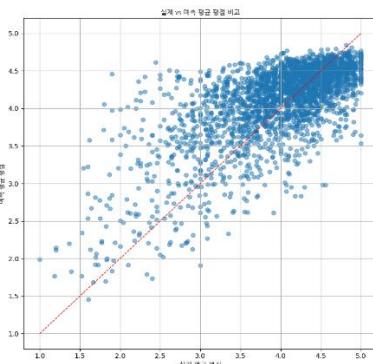
평균 절대 오차 (MAE): 모델은 검증 데이터에 대해 0.3893의 MAE를 기록했다. 이는 1점부터 5점까지의 평점 척도에서 예측 평점이 실제 평점과 평균적으로 약 0.39점의 차이를 보인다는 의미이다. 당초 프로젝트 목표였던 MAE 0.5 이내를 달성하여, 텍스트 정보만으로도 제품의 미래 평점을 예측하는 데 있어 모델이 충분한 정확도를 가짐을 확인할 수 있다.

평균 절대 백분율 오차 (MAPE): MAPE는 **11.19%**로 나타났다. 이는 예측 평점이 실제 평점 대비 평균적으로 약 11.19%의 상대적 오차를 보인다는 의미이다. MAE와 함께 MAPE를 고려할 때, 초기 리뷰 텍스트만으로도 제품의 향후 평점을 상당히 신뢰할 수 있는 수준으로 예측할 수 있음을 나타내며, 이는 신상품 출시 전략 수립 및 초기 시장 반응 예측 등에 유용하게 활용될 수 있다.

4.3. 예측 결과 시각화 및 잔차 분석

모델의 예측 경향과 오차 분포를 시각적으로 확인하기 위해 실제 평점과 예측 평점 비교 산점도 및 잔차 분포 히스토그램을 분석했다.

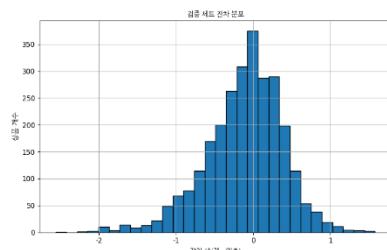
실제 vs 예측 평균 평점 비교



산점도에서 예측값들이 대부분 빨간색 대각선 ($Y=X$) 근처에 밀집해 있음을 알 수 있다. 이는 모델이 실제 평점을 상당히 정확하게 예측하고 있음을 시각적으로 보여준다.

특히 3.5점 이상의 높은 평점대에서는 실제값과 예측값의 일치도가 더욱 높게 나타나, 모델이 긍정적인 평가가 많은 제품군에서 안정적인 예측력을 보임을 알 수 있다.

잔차 분포 분석



검증 세트에 대해 실제 평점과 예측 평점의 차이(잔차)를 계산하여 시각화한 결과, 잔차는 평균 0을 중심으로 좌우 대칭적인 정규분포에 가까운 형태를 보였다. 이는 모델이 전반적으로 특정 평점 구간에 편향되지 않고 안정적으로 예측하고 있음을 시사한다. 특히, 잔차의 대부분이 -0.5에서 0.5 사이에 분포하고 있으며, 이는 예측이 실제 값과 큰 차이를 보이지 않음을 의미한다.

4.4. 예측이 잘 된 사례 분석

모델의 예측 정확도가 높았던 대표적인 사례를 분석하여, 모델이 어떤 기준으로 평점을 예측하는지 해석했다.



사례 텍스트: "does the job pretty well and it leaves a quite refreshing scent unlike the normal ones that make you want to after using for a long time... now my hair seems as healthy as it could be with just this one shampoo"

분석 결과: 이 상품의 실제 평균 평점은 4.43이었고, 모델은 4.32로 예측하여 오차 0.11 수준의 매우 안정적인 결과를 보였다. 텍스트에는 "상쾌한 향", "가족 모두가 좋아함", "모발이 건강해짐" 등 긍정적인 피드백이 주로 포함되어 있었고, 중간에 "눈이 따갑다"는 부정적인 표현도 있었으나 전체 감정 흐름을 왜곡할 정도는 아니었다.

모델 해석: 이 사례는 모델이 단일 리뷰의 특정 문장보다는, 전체 리뷰 텍스트에 담긴 감성적 경향과 긍정/부정 키워드의 빈도를 바탕으로 평균 평점을 안정적으로 예측할 수 있음을 보여준다.

4.5 리뷰 텍스트 해석 방식에 대한 주의점:

본 모델은 단일 리뷰를 기반으로 예측을 수행하는 것이 아니라, 하나의 상품에 달린 복수 사용자의 리뷰를 연결(concatenate) 하여 하나의 텍스트로 구성한 입력을 사용한다. 따라서 모델이 받아들이는 텍스트 입력은 개별 사용자의 감정 흐름이 아니라, 상품 전체에 대한 종합적 평가 텍스트로 볼 수 있다.

이러한 이유로 텍스트 내에는 서로 상반된 긍정적·부정적 표현이 함께 존재할 수 있으며, 문장 간 논리적 일관성보다는 언어적 감정 신호의 분포가 중요한 정보로 작용한다. 모델은 이와 같은 집합적 리뷰 표현을 바탕으로, 해당 상품에 대한 평균 평점을 예측하게 된다.

따라서 결과 해석 시 개별 문장의 흐름보다는, 전체적인 감정 경향과 키워드 사용 분포의 균형을 중심으로 분석하는 것이 적절하다.

이번 프로젝트를 통해, 제품 출시 초기 리뷰 데이터만을 이용하여 향후 평균 평점을 예측하는 딥러닝 모델을 성공적으로 구축하고 검증할 수 있었다. 구조화된 메타데이터 없이 텍스트 데이터만으로도 일정 수준 이상의 예측력이 확보된다는 사실은, 실제 산업 현장에서 데이터 수집이 어려운 상황에도 활용 가능성을 시사한다.

특히 Attention 메커니즘을 통한 문맥 가중 처리 방식은 리뷰 내 핵심 단어를 반영해 모델의 표현력을 향상시키는 데 효과적이었다. 최종적으로 MAPE 11% 수준의 예측 성능을 달성하였으며, 이는 평균 오차가 약 0.39점 이내라는 점에서 실용적으로 의미 있는 결과이다.

5. 결론 및 제언

본 프로젝트는 Amazon 제품의 초기 리뷰 텍스트만을 활용하여 향후 평균 평점을 예측하는 딥러닝 모델을 성공적으로 구축하고 검증하였다. 사용자 정보나 상품 메타데이터와 같은 추가 정보 없이 텍스트 데이터만으로도 일정 수준 이상의 예측력이 확보된다는 사실은, 실제 이커머스 산업 현장에서 신상품 출시 초기의 시장 반응을 빠르게 파악하고 전략을 수립하는 데 유용하게 활용될 수 있음을 시사한다.

모델은 LSTM(Long Short-Term Memory)과 Attention 메커니즘을 결합한 구조로 설계되었다. 특히, 시퀀스의 양방향 정보를 학습하는 양방향 LSTM 구조와 다층 LSTM 레이어를 도입하여 고차원적인 텍스트 패턴을 효과적으로 추출하고자 했다. 각 LSTM 레이어 뒤에 적용된 Dropout은 과적합을 방지하고 모델의 일반화 성능을 높이는 데 기여했다. 또한, Attention 메커니즘을 통해 출력 시퀀스 내 중요한 단어에 더 큰 가중치를 부여함으로써 모델이 핵심 정보에 집중하고 표현력을 향상시키는 데 효과적이었다. 임베딩 레이어를 학습 가능한 구조로 설정하여 리뷰 데이터에 맞는 단어 표현을 스스로 학습하도록 하였으며, 입력 시퀀스 길이를 500으로 제한하여 연산 효율성과 학습 집중도를 모두 확보했다.

최종 모델은 검증 데이터셋에서 평균 절대 오차(MAE) 0.3893 및 **평균 절대 백분율 오차(MAPE) 11.19%**의 예측 성능을 달성했다. 이는 당초 목표했던 MAE 0.5 이내를 충족하며, 평점 척도(1~5점)에서 약 0.39점 이내의 평균 오차를 보이는 실용적으로 의미 있는 결과이다. 잔차 분포 분석 결과, 예측 오차가 평균 0을 중심으로 좌우 대칭적인 정규분포에 가까운 형태를 보였으며, 대부분의 잔차가 -0.5에서 0.5 사이에 분포하여 모델이 전반적으로 편향되지 않고 안정적으로 예측하고 있음을 확인했다. 예측값들이 실제값과 붉은색 대각선에 가깝게 위치한 시각화 결과 또한 모델의 높은 예측 정확도를 뒷받침한다.

이 프로젝트는 초기 리뷰의 감성적·언어적 특징이 이후 평점 변화와 상관관계가 있으며, 소수의 텍스트 데이터만으로도 유의미한 예측 정확도를 달성할 수 있음을 보여주었다. 또한, 임베딩과 LSTM, Attention을 결합한 간단한 구조의 딥러닝 모델만으로도 충분한 학습이 가능하다는 것을 입증했다.

향후 연구에서는 다음과 같은 개선점을 고려해볼 수 있다:

더 긴 시퀀스 처리: 현재 500으로 제한된 시퀀스 길이를 확장하거나, 긴 텍스트를 효과적으로 처리할 수 있는 Transformer 기반 모델을 탐색하여 리뷰 텍스트 후반부 정보 손실에 대한 영향을 줄일 수 있다.

과적합 방지 강화: Early Stopping, 더 다양한 정규화 기법 또는 데이터 증강(Data Augmentation)을 적용하여 모델의 일반화 성능을 더욱 향상시킬 수 있다.

평점대별 성능 분석: 낮은 평점대 예측에 대한 모델의 성능을 보완하기 위해, 평점 불균형 문제를 해결하기 위한 샘플링 기법(예: 오버샘플링, 언더샘플링) 또는 손실 함수(Loss Function) 조정을 시도해볼 수 있다.

외부 데이터 활용: 사용자 정보, 상품 카테고리, 브랜드 정보 등 텍스트 외의 메타데이터를 통합하여 예측 모델의 성능을 추가적으로 개선할 가능성이 있다.

이러한 후속 연구를 통해 본 모델의 예측력을 더욱 고도화하고, 이커머스 분야에서 실질적인 비즈니스 가치를 창출할 수 있을 것으로 기대된다.

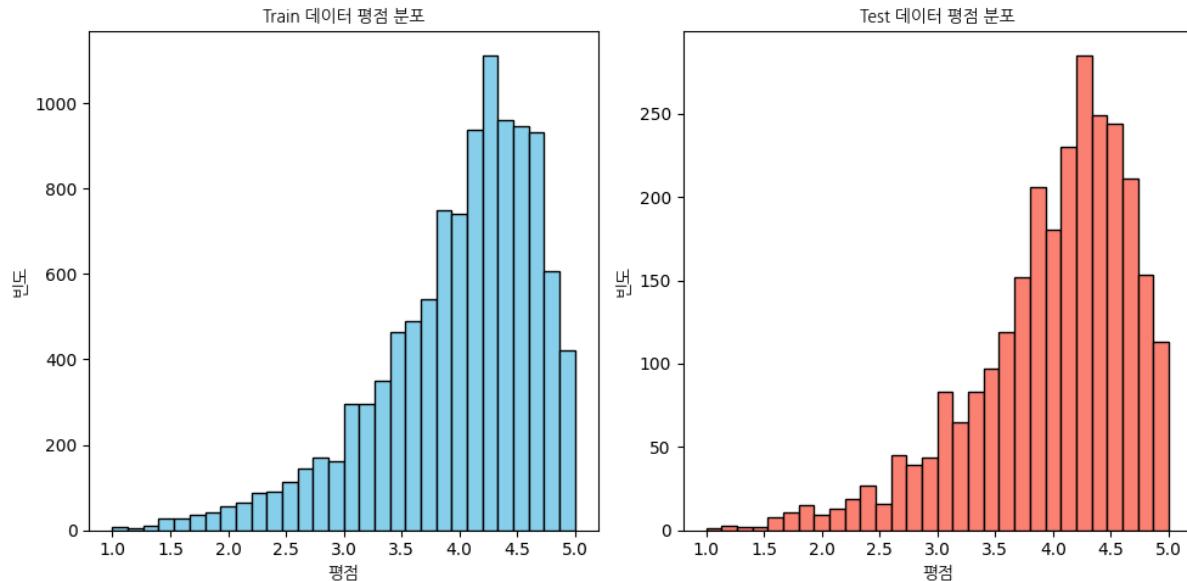
6. 참조(Reference)

Data sets: <https://www.kaggle.com/datasets/wajahat1064/amazon-reviews-data-2023> (This is a large-scale Amazon Reviews dataset, collected in 2023 by McAuley Lab)

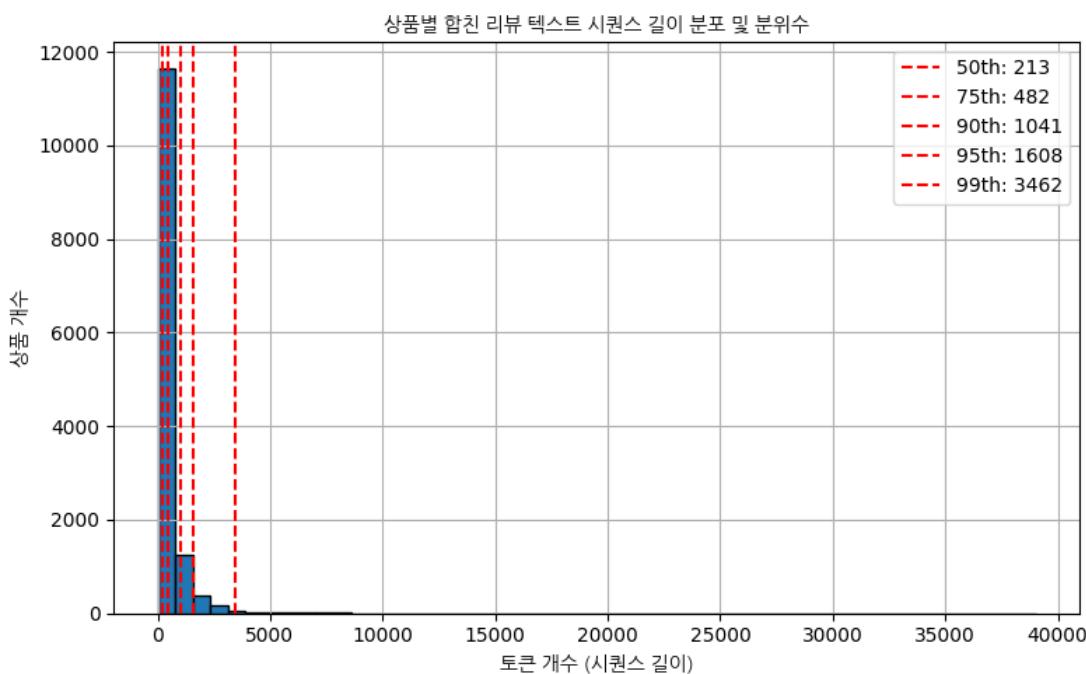
참고 자료: <https://wikidocs.net/22893> (wikidocs about attention mechanism)

<https://wikidocs.net/182469> (wikidocs about keras tokenizer)

train/test set rating distribution:



Length of sequence:



오차 0.3이하인 샘플:

		리뷰 텍스트	실제 평점	예측 평점	오차
3		so i had this awesome vintage necklace after w...	3.818182	4.027779	0.209597
4		good <OOV> i must have gotten a lemon it took ...	4.264706	4.011184	0.253522
12		i am 2 weeks in and people have told me they've...	4.000000	3.864112	0.135888
19		seems like a pretty good quality of hair very ...	4.111111	4.400019	0.288908
25		does the job pretty well and it leaves a quite...	4.428571	4.316956	0.111616