기계 학습을 통한

혐오 표현 분류

B677011 방찬웅

목차

1. **서론**[**3**](#three)

1-1 온라인 세상의 혐오[3](#three)

1-2 혐오 표현의 사례[3](#three)

1-3 게임에서의 악성 채팅[3](#three)

1-4 혐오 표현 분류[4](#four)

1-5 목표[4](#four)

1. **본론**[5](#five)

2-1 데이터 수집[5](#five)

2-2 데이터 분석[5](#five)

2-3 모델링 학습[10](#ten)

2-4 첫 번째 테스트[12](#twelve)

2-5 두 번째 테스트[15](#fifteen)

2-6 세 번째 테스트[16](#sixteen)

1. **결론**[**18**](#eighteen)

3-1 테스트 결과 비교[18](#eighteen)

3-2 문제점 및 한계[20](#twenty)

3-3 이 후 과제[21](#twentyone)

1. **서론**

1-1 온라인 세상의 혐오  
스마트폰의 대중적인 보급과 인터넷이 확산됨에 따라, 남녀노소 불문하고 온라인 세상과 소통할 수 있는 시대가 되었다. Facebook, Instagram을 비롯한 SNS, 동영상 플랫폼인 TikTok, YouTube등 다양한 플랫폼들이 인기를 끌면서 연예인, 일반인 구분없이 많은 인기를 얻게 되고, 소통하는 시대가 되었다. 하지만 인터넷 매체의 증가에 따라 댓글이 폭발적으로 증가하고, 온라인 채팅 시스템이 많아지면서 온라인 상의 혐오 표현이 많아 짐에 따라 심각한 사회적 문제가 되었다. 익명성이 보장되고 서로 얼굴을 볼 수 없는 온라인이라는 특성을 악용하여 혐오 적인 표현을 아무렇지 않게 표현할 수 있는 문제가 생겼다. 특히 한국에선 집단과 집단과의 갈등을 일으키는 더욱 심각한 현상으로 자리 잡고 있다.

* 1. 혐오 표현의 사례  
     악성 댓글로 인해 연예인, 인플루언서 등 많은 이들이 고통과 피해를 받고 있다. 이에 시달린 [[1]](#footnote-1)유튜버와 배구선수가 자살을 하는 충격적인 사건이 벌어졌다. 또한 [[2]](#footnote-2)연예인의 자살 사건과 악성 댓글의 관계에 대한 통계에서 영향을 미친다는 응답자가 98%에 육박할 정도로 심각한 수준이다. 이처럼 악성 댓글은 끔찍한 결과를 초래할 수 있는 문제이다. 또한 [[3]](#footnote-3)특정 집단에 대해 혐오 표현을 사용하여 비난하거나 비하하는 문제도 심각하다.
  2. 게임에서의 악성 채팅  
     웹 사이트, 플랫폼 에서의 혐오 표현도 심각하지만 실시간으로 채팅을 주고받는 [[4]](#footnote-4)게임 내에서도 혐오 표현이 많이 이루어지고 있다. 게임 내에서의 욕설은 일상처럼 일어나고 있으며 욕설뿐 아니라 가족을 욕하는 ‘패드립’, 여성 혐오와 인종 차별까지 자주 일어나고 있다. 심지어 게임에서 싸움을 하다 현실로 이어져 폭력으로 이어지는 ‘현피’까지 이어질 수 있다. 익명성이 보장되고 주 소비 연령층이 낮기 때문에 혐오 표현의 사용이 더 잦게 일어난다.
  3. 혐오 표현 분류  
     현재 깨끗한 온라인 세상을 만들기 위해 정부, 기업 등 많은 노력을 하고 있다. 처벌 수위를 높이거나, 댓글 시스템을 없애는 방법도 시행되고 있다. 하지만 애초에 혐오 표현이 들어간 글을 사전에 감지하여, 이를 막는 것이 가장 좋다고 생각한다. 좋은 글은 남기되 나쁜 글은 없애자는 취지로 이 프로젝트를 진행하게 되었다.
  4. 목표  
     기계 학습 방법을 활용해 입력으로 문장을 받아 이를 분류하는 기능을 만드는 것이 목표이다. 또한 악성, 혐오 표현으로 감지된 문장이 어떤 종류의 혐오 표현인지 결과를 수집하여 들어온 데이터 중 어떤 종류의 혐오가 많이 발생하고 있는지 알아본다.

1. **본론**

2-1 데이터 수집  
사용한 데이터는 [[5]](#footnote-5)스마일게이트 AI센터에서 제공하는 악플 및 혐오 발언 데이터셋 ‘UnSmile’ 데이터셋을 사용한다. 이 데이터셋에서의 혐오 표현은 “특정 사회적(소수자) 집단에 대한 적대적 발언, 조롱, 희화화, 편견을 재생산하는 표현”으로 정의하였다. 혐오 표현이 포함되어 있지 않은 일반적인 악플이나 욕설은 혐오 표현으로 분류하지 않고 악플/욕설로 분류한다. 데이터 개수는 총 18,742개이다.

2-2데이터 분석  
데이터셋은 huggingface datasets 형식으로 되어있으며 Train, Valid 데이터 셋이Dictionary 형태로 되어있고, 그 안에 하나의Dictionary가 더 들어가 있는 구조이다. Train 데이터셋 15005개, Valid 데이터셋 3737개로 구성 되어있다. 테이블이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

문장과 문장의 종류, 어떤 종류의 문장인지에 대한 레이블로 구성 되어있다.

데이터셋 안에 들어있는 혐오 표현 종류들의 각각의 개수를 알아보기 위하여 표를 만들고자 하였다. 따라서 현재 데이터셋에서 정보를 추출해 Pandas 형식으로 새로운 데이터 프레임을 만들었다. Train 데이터 셋에서 열의 Label 값이 1인 개수를 추출하여 통계를 내어 표를 만들었다.

텍스트이(가) 표시된 사진

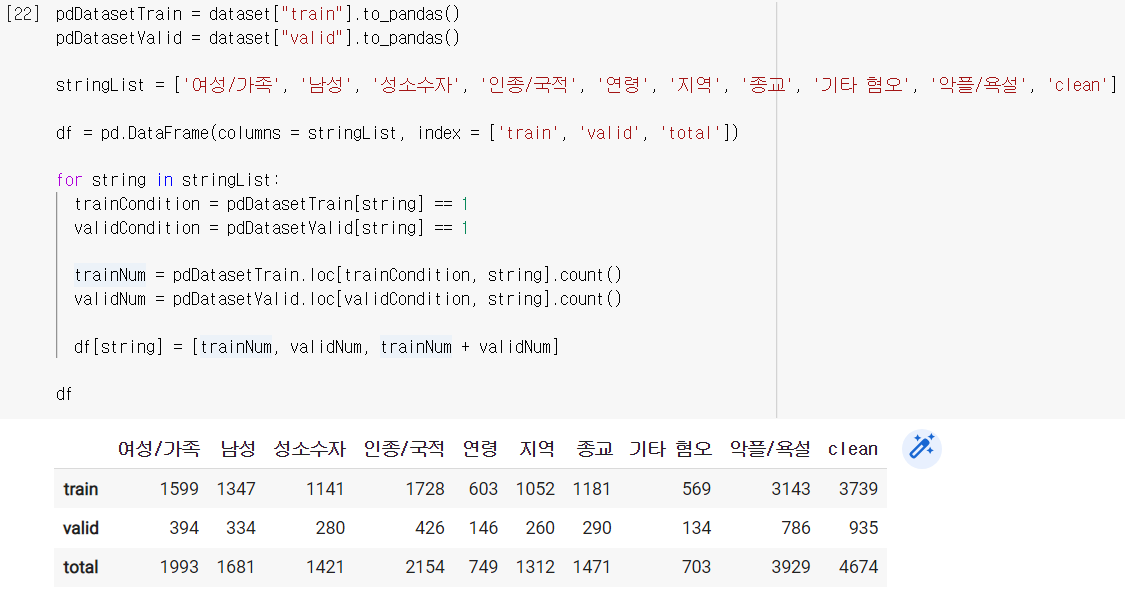
자동 생성된 설명

표를 만든 뒤, 코드 상 공통적인 부분이 많기 때문에 코드를 효율적으로 줄이기 위하여 혐오 표현 종류를 문자열 List로 만들어 다시 작성하였다.

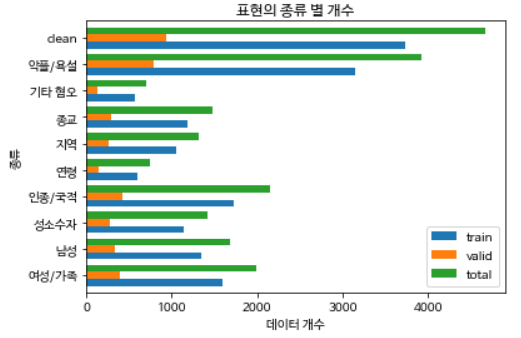
텍스트이(가) 표시된 사진

자동 생성된 설명

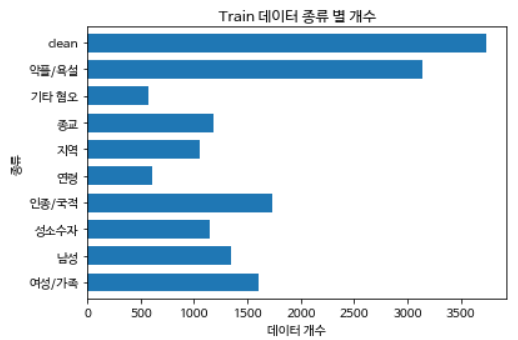
같은 결과로 잘 작동 하였다. Valid 데이터셋과 두 데이터셋을 합친 Total항목도 추가하였다.

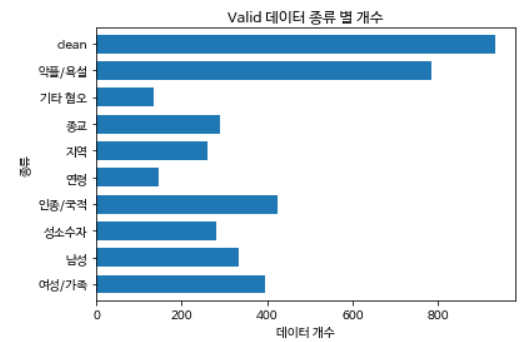


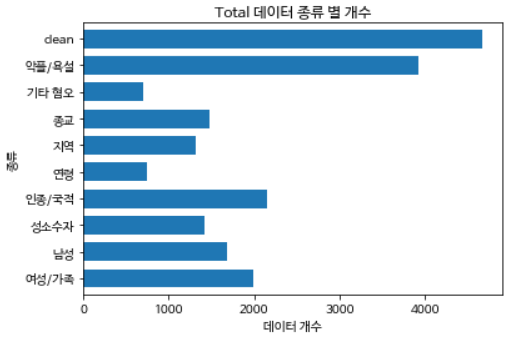
더욱 직관적으로 보기 위해 Matplotlib를 이용하여 Bar 그래프를 그렸다.



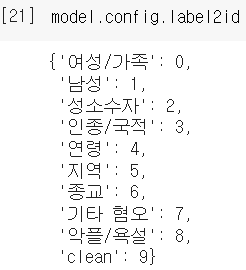
각각 데이터의 비교가 어렵기 때문에 Train, Valid, Total을 분리하여 다시 막대 그래프를 그렸다.



Train 데이터 셋에서 가장 많은 3개의 데이터(Clean과 악플/욕설은 혐오 표현이 아니므로 제외)는 인종/국적, 여성/가족, 남성이며, 가장 적은 3개의 데이터는 기타 혐오, 연령, 지역이다. 

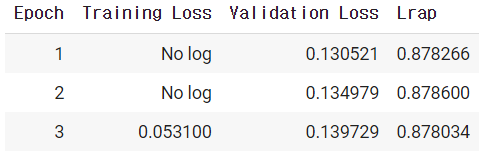
Valid 데이터 셋에서 가장 많은 3개의 데이터는 인종/국적, 여성/가족, 남성이며, 가장 적은 데이터 3개는 기타 혐오, 연령 지역이다. Train 데이터와 Valid 데이터를 비교한 그래프의 구조가 굉장히 유사한 걸로 보아 데이터 셋의 데이터 비율이 비슷한 비율로 구성되어 있음을 알 수 있다. 

전체 데이터를 비교한 그래프 역시 유사한 비율로 보여진다.

2-3 모델링 학습  
다중 레이블 분류 모델을 사용하여 혐오 표현 종류 마다 점수를 매기고 가장 높은 점수의 종류를 결과로 본다. Activation Function으론 Sigmoid 함수를 사용한다. 학습은 총 세가지를 비교할 예정이다.  


Learning rate = 0.00002, Batch size : 64, epoch : 5 로 학습 했을 때의 Loss이다.   
테이블이(가) 표시된 사진

자동 생성된 설명

Learning rate = 0.00001, Batch size : 64, epoch : 3 로 학습 했을 때의 Loss이다.   


Learning rate = 0.00001, Batch size : 64, epoch : 5 로 학습 했을 때의 Loss이다.   
테이블이(가) 표시된 사진

자동 생성된 설명  
설정한 Learning Rate와 epoch에 따라 손실률이 차이가 나는 것을 확인할 수 있다. 이 세가지 방법으로 테스트를 진행해 비교 해보겠다.

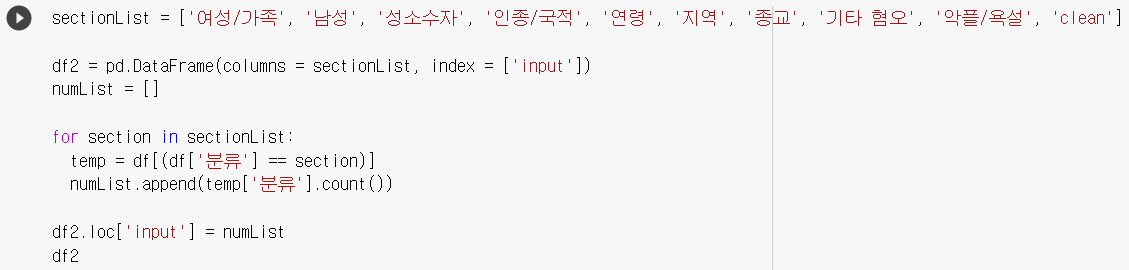
2-4 첫 번째 테스트  
모델 테스트를 위해 [[6]](#footnote-6)kocohub에서 제공하는 Korean HateSpeech 데이터 셋을 불러왔다.  
  
불러온 데이터 셋으로 분류를 실행했다.  
텍스트이(가) 표시된 사진

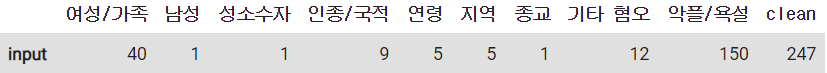
자동 생성된 설명

문장 각각의 혐오 표현 종류의 대한 점수를 산출하고 가장 높은 점수와 종류를 데이터 프레임으로 만들어 시각화했다.



분류한 데이터 셋의 혐오 표현 종류 별 개수를 종합했다.



  
이 데이터 셋엔 여성/가족, 인종/국적에 대한 혐오 표현이 많은 걸 확인할 수 있다.

이 중 점수가 낮은 데이터(0.6 이하)를 추출해보았다. 총 개수는 26개 였으며 점수들의 평균은 0.412946이다. 점수가 낮은 데이터는 특정 인물, 대상을 지칭하는 고유 명사를 사용하거나, 딱히 가르키는 대상이 없는 경우 점수가 낮았다.

텍스트이(가) 표시된 사진

자동 생성된 설명

이번엔 점수가 높은 데이터(0.95 이상)를 추출해보았다. 총 개수는 355개 였으며 점수들의 평균은 0.986845이다. 점수가 높은 데이터는 대부분 Clean, 악플/욕설 데이터였다. 혐오나 욕설이 없는 표현은 높은 확률로 Clean이 되고 욕설이 있는 문장은 악플/욕설로 많이 분류되었다. 또한 특정 집단을 나타내는 단어(예: 맘충, 여자, 전라도 등)가 포함된 문장은 점수가 높게 분류되었다.



첫 번째 테스트한 방법의 성능 평가이다.

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

\* 0 ~ 9  
여성/가족, 남성, 성소수자, 인종/국적, 연령, 지역, 종교, 기타 혐오, 악플/욕설, clean

2-5 두 번째 테스트

두 번째 방법로 같은 데이터 셋으로 테스트를 진행 하였다.



점수가 낮은 데이터(0.6 이하)를 추출해보았다. 총 개수는 40개 였으며 점수들의 평균은 0.451837이다.

텍스트이(가) 표시된 사진

자동 생성된 설명

점수가 높은 데이터(0.95 이상)를 추출해보았다. 총 개수는 257개 였으며 점수들의 평균은 0.97112이다.



두 번째 테스트한 방법의 성능 평가이다.

테이블이(가) 표시된 사진

자동 생성된 설명

2-6 세 번째 테스트

세 번째 방법도 같은 데이터 셋으로 테스트를 진행 하였다.



점수가 낮은 데이터(0.6 이하)를 추출해보았다. 총 개수는 26개 였으며 점수들의 평균은 0.368803이다.



점수가 높은 데이터(0.95 이상)를 추출해보았다. 총 개수는 372개 였으며 점수들의 평균은 0.990772이다.

테이블이(가) 표시된 사진

자동 생성된 설명

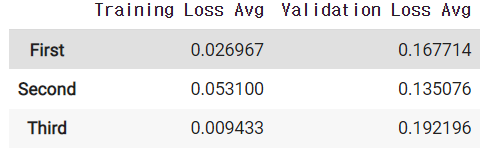
세 번째 테스트한 방법의 성능 평가이다. 테이블이(가) 표시된 사진

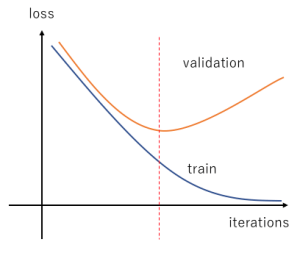
자동 생성된 설명

1. **결론**

3-1 테스트 결과 비교

학습 시 발생한 Training Loss와 Validation Loss를 평균내어 비교 해보았다.

  
모든 테스트의 Training Loss는 Validation Loss보다 낮았다. 세 번째 시도의 Training Loss의 평균이 현저히 낮았다. Learning Rate을 낮추고, Epoch를 늘린 효과이다. 하지만 전체적으로 모든 테스트에서 Training Loss가 감소함에 따라 Validation Loss는 증가하는 현상이 나왔다. [[7]](#footnote-7)과적합이 발생한 것 같다.



다음은 데이터 셋의 혐오 표현 종류 별 개수를 비교해보았다.

  
Learning Rate를 낮춘 테스트에서 clean 데이터가 많은 증가를 하였다. 첫 번째 테스트에서 혐오, 욕설로 구분된 문장들이 clean으로 많이 분류되었다.

다음은 점수가 낮은 데이터(0.6 이하)의 개수와 점수 평균을 비교해보았다.

텍스트, 테이블이(가) 표시된 사진

자동 생성된 설명  
epoch가 5인 테스트에서 개수는 26개로 같았다. Learning Rate가 낮고 epoch가 큰 테스트에서 평균 점수가 가장 낮았다. 평균 점수가 낮을수록 알 수 없는 문장에 대한 분류가 잘 이루어 지지 않았다고 생각한다.

다음은 점수가 높은 데이터(0.95 이상)의 개수와 점수 평균을 비교해보았다.

텍스트, 테이블이(가) 표시된 사진

자동 생성된 설명  
epoch가 5인 테스트에서 개수가 350개 이상으로 많았고, 점수가 0.985이상으로 높았다. 점수가 높았던 분류는 대부분 clean 데이터였다.

다음은 성능 평가를 비교해보았다. [[8]](#footnote-8)성능 평가 비교에는 다중 레이블 분류에 사용되는 Samples Average 비교하였다.

테이블이(가) 표시된 사진

자동 생성된 설명  
점수들을 보았을 때 모두 점수는 준수한 것을 보여주는 것을 확인할 수 있다.

비교한 내용으로 종합해 보았을 때 첫 번째 테스트에서 진행한 방법이 가장 분류에 적합하다고 생각한다. 혐오 표현 분류의 목적은 깨끗한 온라인 세상을 만드는 것이 목적인데 좋지 않은 표현이 포함되어 있는 문장이라면 필터링 할 수 있어야 한다고 생각한다. 다른 방법보다 애매한 문장에 대해 clean으로 분류하지 않고 다른 종류로 분류를 하고, 다양하게 분류하면서도 점수가 높은 데이터의 수가 많고 평균 점수가 높았던 첫 번째 테스트의 방법이 가장 프로젝트에 적합한 방법이다. Learning Rate가 낮다고 해서 무조건 좋은 것은 아닌 것 같고, epoch는 충분하게 설정해야 좋을 것 같다.

3-2 문제점 및 한계

문장이 길어지고 다양한 단어가 나오는 문장에 대한 분류가 잘 이루어지지 않았다. “전현무 넘싫음..여친이 야구선수만 주구장창 만나왔으니..본인체력 딸려서 개고기먹는 전현무극혐..미식회나와서 개고기극찬..노견 또또대하던 태도” 라는 문장의 결과는 0.37의 점수로 성소수자로 분류 되었다. 이처럼 인물, 성별, 동물, 악플 종합된 문장은 제대로 분류하지 못하였다.

특정 인물, 특정 대상을 지칭하는 고유명사가 포함되고 심한 욕설이 아닌 비난, 비하 문장에 대한 분류도 잘 이루어지지 않았다. 이런 문장들이 사회적으로 좋지 않은 문장이지만 그렇다고 혐오, 욕설로 분류되기엔 애매했던 문장이다. 이런 문장을 clean이나 엉뚱한 종류로 분류를 하였다.

데이터 셋을 살펴본 결과 레이블이 주관적이라는 한계가 있다. 본인이 볼 때 절대 이 레이블이 아닐 것 같은 문장들이 있었다. 예를 들어 “적화 다 된거냐?ㅋㅋㅋㅋㅋㅋ 기가 차다, 빨갱이들이 많긴 많은 가봐 이나라에” 란 문장의 레이블은 기타 혐오인데, 이 문장은 인종/국적에 레이블이 되어야 맞다고 생각한다. 이와 같은 레이블링이 성능에 분명이 영향이 있었을 것이라고 본다.

이와 같이 어려운 문장을 제외하면 충분히 일반적인 혐오 표현 분류를 성공적으로 해주었고, 누가 보기에도 불편한 문장들은 필터링 해줄 수 있다고 본다. 또한 만약 테스트한 데이터 셋이 한 커뮤니티의 댓글 모음이라고 생각한다면, 그 커뮤니티에선 현재 어떤 혐오 표현의 정도가 가장 심각한지 알아볼 수 있고 통계를 내볼 수 있다. 이 정도의 필터링 기능만 해주더라도 온라인 세상이 더욱 깨끗해질 것이라고 생각한다. 더 많고 충분한 데이터와 합리적인 레이블이 되어있는 데이터 셋을 사용한다면 성능이 더욱 좋아질 것이라고 생각한다.

3-3 이 후 과제  
고유명사를 인식할 수 있도록 학습하여 AI가 몰랐던 특정 대상에 대한 이해를 하여 분류 성능이 증가할 것이라 예측한다.

과적합이 발생했다고 생각되기 때문에 데이터 셋의 양을 늘리고, Train 데이터와 Valid 데이터를 적절히 조절하여 과적합 현상을 해결해야한다. 현재는 데이터가 한정되어 있기 때문에 현상을 해결하긴 어렵다.

성능 평가 지표에 대한 분석이 어려워 이번 프로젝트에 가장 적합하고 정확한 정보를 줄 수 있는 것이 어떤 것인지 판단하지 못하였다. 대부분 높은 점수를 기록했지만 더욱 정확한 성능 평가를 위해선 알아야 할 것 같다.

1. 악성 댓글로 인한 자살 사례  
   https://imnews.imbc.com/replay/2022/nwtoday/article/6339502\_35752.html [↑](#footnote-ref-1)
2. 악플과 자살과의 관계 통계  
   http://www.knnews.co.kr/news/articleView.php?idxno=1314618 [↑](#footnote-ref-2)
3. 특정 집단 혐오  
   https://www.snpo.kr/bbs/board.php?bo\_table=npo\_aca&wr\_id=4683 [↑](#footnote-ref-3)
4. 게임 속 혐오와 차별  
   https://www.nocutnews.co.kr/news/5149487 [↑](#footnote-ref-4)
5. UnSmile 데이터셋  
   https://smilegate.ai/2022/03/28/unsmile-dataset/ [↑](#footnote-ref-5)
6. 한글 혐오 표현 데이터 셋  
   https://github.com/kocohub/korean-hate-speech [↑](#footnote-ref-6)
7. Learning Curves, Overfitting  
   https://www.baeldung.com/cs/learning-curve-ml [↑](#footnote-ref-7)
8. Multi Label Model Evaluation  
   https://www.kaggle.com/code/kmkarakaya/multi-label-model-evaluation [↑](#footnote-ref-8)