# 한국어 욕설 채팅 분류 분석

권혁민, 김도현, 이정우
인공지능융합전공, 국어국문학과,
소비자학과
성균관대학교
*2024312676, 2023314598, 2021310896*

***Abstract*—욕설은 상대방의 불쾌감을 유발하며, 특히 온라인 공간에서의 욕설은 그 공격성이 강해지고 욕설 사용자의 심적 부담이 덜하다는 점에서 사용에 주의를 필요로 한다. 욕설 자동 필터링 시스템이 개발되고 사용되어 왔지만, 실제 인터넷과 게임 유저들이 체감하기에는 의문을 가질 부분이 많다. 한국어의 특성에 의해 욕설은 다양한 변이형을 가지며 이에 따라 기존의 욕설 분류 시스템은 정확성, 변형어에 대한 대응을 따라가기 위한 노력이 더욱 필요하다. 욕설 분류를 위한 선행 연구에서는 대부분 딥러닝에 기반한 모델을 사용했다는 점을 바탕으로, 최적의 머신러닝을 구현하는 것을 목표로 설정했다. 이에 따라 본 프로젝트는 TF-IDF, fastText, doc2vec 의 임베딩 기법과 SVM, 로지스틱 회귀의 특성을 이해하고 조합을 구성하여 성능을 비교하였다. 비교 후 벡터의 특성에 따라 SVM 에 XGBoost 모델을 추가하여 최종 모델 구현하였고 최종적으로 약 83%의 정확도를 얻었다. 비교를 위해 간소화한 딥러닝에서 90%의 정확도를 얻기도 하였으며 본 연구를 바탕으로 추후 욕설이 사용되는 상황과 문맥을 고려하여 정확도 높은 모델 연구를 제안한다.**

***Keywords*—*욕설 분류, 온라인 채팅, 머신러닝***

## I. INTRODUCTION

사람과 사람 사이의 의사소통은 다양한 방식으로 이루어진다. 오늘날 우리 사회의 의사소통은 직접 대면하여 이루어지기보다 비대면으로 진행되는 경우가 많다. 인터넷의 발달과 스마트폰의 대중화로 대부분의 사람들은 카카오톡이나 문자 메시지를 통해 대화하고, SNS 를 활용하여 서로의 일상을 공유하며, 온라인 커뮤니티 속에서 인간 관계를 형성하기도 한다.

점점 확장되는 온라인 공간 속에서 사람들은 서로의 의견을 공유하고 유대감을 느낀다. 그러나 이 과정에서 사람들은 좋은 말만 뱉지 않는다. 온라인 공간의 가장 큰 특성인 익명성을 바탕으로 사람들은 더욱 공격적이고 심한 희롱을 서슴지 않는다. 실제로는 서로 만나본 적이 없는 사람에게도 익명이라는 조건 하나로 상대에게 해서는 안 될 말을 한다.

이러한 점에서 온라인 소통 사이의 욕설은 오프라인 공간에서의 욕설과는 다른 영향력을 가진다. 흔히 온라인 공간은 개인적인 공간, 사적인 영역으로 간주한다. 개개인이 원하는 시간에 본인의 의사표현을 환경에 제약 없이 가능하게 하는 공간이기 때문이다. 그러나 온라인 공간의 중요한 특성인 빠른 정보 전달은 사적인 의견을 개인의 영역을 넘어서 더 큰 영향을 미칠 수 있다. 시작은 한 명의 말일지라도 빠르게 전달되고 확산되어 거대한 집단의 말이 되어 공적인 힘을 가지게 될 수 있기 때문이다. 인터넷 뉴스의 댓글, 인터넷 커뮤니티, 게임에서의 채팅 속에서 사람들이 주고받는 말들은 이러한 특성과 영향력을 보여주는 중요한 매개체이다[1].

뉴스 댓글의 경우 흔히 '악플'이라 칭하는 욕설과 비난 댓글은 기사의 당사자 혹은 독자로 하여금 불편한 감정을 유발한다. 지난 2020 년 포털 3 사의 연예 댓글 폐지 또한 이러한 영향을 받아 이루어졌다. 댓글 시스템의 확산에 따라 거짓 뉴스, 이에 대한 비난 댓글이 퍼지는 속도도 빨라졌기 때문이다. 유명인들은 이를 비롯한 이유들로 공황장애 등 심적인 부담을 겪고 심한 경우 자살로 이어지기도 한다.

이는 온라인 커뮤니티의 영향도 적지 않으며 뉴스의 댓글에 비해 욕설의 사용에 제약이 덜한 커뮤니티 공간에서는 서로 욕설을 나누며 더욱 강한 네트워크를 형성한다. 일간베스트와 같이 욕설을 많이 사용하는 커뮤니티의 경우 이러한 경향이 더욱 강하게 나타나기도 한다[2]. 이 과정에서 오고 가는 말들은 다시 기사화되어 또다른 소통의 장이 열릴 수 있고 끊임없이 당사자를 괴롭힐 수 있기 때문이다.

게임 채팅의 경우에는 온라인 게임의 팀원 또는 다른 유저와의 다툼으로 욕설이 사용되는 경우가 많으며 상대의 부모를 욕하는 경우, 인신 공격 등 다양한 방식으로 비속어를 사용한다. 넥슨 인텔리전스랩스 인게이지먼트 연구팀에 따르면 유저들 중 게임 플레이 과정에서 욕설로 인해 불쾌한 경험을 겪은 경우 플레이 시간은 평균 84 분 감소한다. 욕설 이외의 사유로 불쾌한 경험을 겪은 유저의 경우 평균 29 분 감소하는 것과 큰 차이를 보였다. 재접속률도 비속어의 노출이 잦을수록 더 떨어지는 것으로 나타났다[3].

게임 채팅과 댓글 시스템에는 욕설 및 비속어를 자동적으로 판별하여 가리거나 삭제하는 기능이 존재한다. KISO 가 개발한 이용자보호시스템(KSS)의 경우 1 년 간 33 만 개의 욕설을 걸러내었다고 한다. 그럼에도 불구하고 온라인 공간에서의 욕설은 다양한 변이형을 가지며 창의적인 형태로 사용된다. KSS 가 이 부분을 어느정도 해결해가는 중이지만 욕설을 통제한다 해도 계속해서 새로운 방식으로 욕설을 사용하고 그 과정에서 유대감을 형성하는 사회적인 기능을 막지 못한다는 것이다. 그 사회적 기능을 무시할 수 없지만, 이로 인한 문제점들은 엄연히 해결해야 할 것이다[4].

더하여, 최근에는 생성형 AI 의 기술이 급속도로 성장하여 AI 가 만들어내는 문장을 마주하는 경우도 적지 않다. 생성형 AI 를 학습, 연구의 목적으로 활용하여 작업의 효율성을 높이는 것은 기술 발전의 순기능이다. 그러나 최근 AI 기업들은 이를 다른 방식으로 활용하여 경제적 이득을 취하려 하기도 한다. 업계는 그동안 성적인 내용, 정치적으로 민감한

내용에는 답변을 내놓지 않도록 개발되었다. 그러나 이러한 기준을 없애 성적 대화가 가능하게 하는 성인용 챗봇을 개발하여, 기존에는 불가능한 대화를 가능케 했다. 이를 남용한다면 꽤나 큰 사회적 문제를 일으킬 수 있다. AI 가 성적인 비속어를 학습하여 사용한다면, 현재 온라인 상에서 오고 가는 욕설이 가지는 문제 이상으로 큰 파장을 가져올 것이다[5].

본 연구는 이러한 온라인 공간에서의 욕설이 가지는 문제들을 바탕으로 알고리즘이 욕설을 분류하는 과정에 대해 알아보고자 하였다. 욕설을 분류하는 시스템이 현재도 존재하긴 하지만, 앞서 설명하였듯 새로운 형태로 비속어를 사용하는 흐름을 따라가기에는 지속적인 연구와 개발이 필요하다. 또한 욕설이 아닌데 욕설이라 판단하는 경우, 욕설을 제대로 분류하지 못하는 경우도 아직까지 존재한다. 기존의 시스템이 적지 않은 효과를 가지고 있음은 분명하나 이용자의 입장에서 이를 체감하기에 아직 부족하다고 판단했다.

한글은 어근과 접사에 따라 의미를 가지며 조사 결합이라는 특징을 가진 교착어이다. 온라인 공간 속에서는 다른 언어에 비해 어순이 강제되지 않고, 띄어쓰기도 잘 지켜지지 않으며 이런 이유로 연구에 어려움이 있는 언어이다. 꾸준한 연구가 필요한 언어라 이해할 수 있다. 따라서 욕설이 어떻게 분류되는지에 대해 알고, 수업을 통해 배웠던 여러 모델들을 통해 어떻게 높은 정확도 높은 모델을 만들어낼 수 있을지 파악하고자 했다. 배운 범위 내, 약간의 조사를 바탕으로 현재의 기술에 대해 이해하고, 추후 연구를 통해 상황, 문맥에 따른 욕설 분류가 가능하도록 하여 온라인 상에서 남용되는 비속어를 더욱 효과적으로 통제할 수 있을 것이라는 기대에 다음과 같은 연구를 진행하였다.

## II. METHOD

욕설 분류 모델의 경우 이미 적지 않은 선행 연구가 이루어졌다. 따라서 연구를 위해서는 자료들과 알고 있는 지식을 바탕으로 우리만의 연구를 진행하기 위한 방법이 필요했다. 온라인 공간에서 사용되는 욕설은 정확한 욕설을 단어 그대로 사용하는 경우와 신조어나 욕설에 기호 등을 더하여 욕설이 아닌 것처럼 표현하는 경우(‘^^lqkf’, ‘시@발’)가 존재한다. Kaggle 을 통해 한국어 채팅 욕설 분류 데이터를 찾았고 500 만 개의 충분한 데이터가 있었으며 욕설 그 자체가 포함되어 있는 문장, 변이형이 포함된 문장이 모두 충분히 있었기 때문에 해당 데이터를 사용하였다.

현재 업계에서 사용되고 있는 욕설 분류 시스템, 기존의 연구에서 사용된 모델은 대부분 딥러닝을 기반으로 하였다. N-Gram, Lexicon 등을 활용하여 지도학습, 비지도학습을 구현한 머신러닝 연구도 있었다[6]. 그러나 해당 연구가 진행되었을 때에 비해 다양한 변형어가 만들어졌고 딥러닝에 비해 꾸준한 연구가 이루어지지 않았다는 점에 주목하였다.

딥러닝에 기반한 욕설 분류 모델의 경우, 게임 채팅에서의 욕설 탐지 연구[7]에서는 딥러닝의 콘볼루션 신경망을 사용하여 모델을 만들었으며, HSV 스케일 이미지 처리 연구[8]에서도 HSV 스케일로 인코딩한 뒤 Fast R-CNN, Mask R-CNN, FPN(Feature Pyramid Network)을 기반으로 한 모델을 사용하여 학습하였다.

그러나 주어진 환경에서 딥러닝을 구현하기에는 시공간적 제약이 존재한다. 딥러닝을 통해 모델을 학습시키기 위해서는 많은 데이터와 시간이 필요하며 이를 비롯한 많은 연산 비용을 필요로 한다. 반면 머신러닝은 이러한 딥러닝의 단점을 해결할 수 있는 방법이 될 수 있다. 학기 중 배운 내용 또한 머신러닝이 큰 비중을 차지하기 때문에 머신러닝을 활용하여 최적의 모델을 만들어내는 것을 최우선하여 프로젝트를 진행하였다. 더불어 딥러닝이 과연 욕설 분류에 있어 정말 높은 성능을 보이는지 검증하고, 프로젝트에 딥러닝을 시도해본다면 향후 연구의 기초가 될 것이라 판단하여 데이터를 간소화하여 모델을 구현해보았다.

우선, 주어진 한글 데이터를 임베딩하기 위한 방법으로는 여러가지가 있다. 최적의 모델을 찾아가기 위해서는 여러 방법을 시도해보고 비교해보는 과정이 필요하다. 이를 위해 수업 시간에 배운 TF-IDF 와 fastText, doc2vec 기법을 활용하여 임베딩을 진행하였다.

TF-IDF 는 문서의 집합에서 중요도를 계산하기 위해 빈도와 역문서 빈도를 고려한다. 이 과정에서 각각의 단어를 토큰화한다는 특징이 있으므로 정확한 욕설에 대한 구별이 가능하다는 장점이 있다. 반면 fastText 기법은 단어를 쪼개서 훈련한다는 특징이 있다. 이는 조사와 어미가 발달한 한글 데이터에 적합하고 은어, 신조어, 변형어를 분류하는 데에 강점이 있다. 이는 욕설의 변이형을 분류하기에 적합하다. doc2vec 은 다량의 코퍼스를 임베딩하는 경우와 문서 간의 유사성을 파악하고 분류하는 데에 좋은 성능을 보이는 기법이다.

위 방법으로 임베딩을 진행하여 각각에 대해 머신러닝 모델을 구현해보았고, 최종 모델에 사용하기 위해서는 TF-IDF 와 fastText 를 병합하여 사용하였다. 욕설 그 자체가 포함된 데이터와 단어를 다른 방식으로 표현한 데이터를 정확하게 분류하는 데에 강점이 있기 때문이다.

임베딩한 데이터를 가지고 우수한 학습을 진행시키기 위해서는 어떤 모델을 선정하는지도 중요한 과정 중 하나이다. GridSearch 는 모든 하이퍼파라미터에 대해 최적의 조건을 찾아 해당 모델의 하이퍼 파라미터 조합을 찾아내는 모델이다. 머신러닝을 수행하며 과대적합이 일어나지 않아야 하므로 최적의 규제 값을 찾아내기 위해 GridSearch 를 기반으로 한 로지스틱 회귀 모델을 설정하였다. fastText 와 TF-IDF 를 병합하여 임베딩한 데이터는 일반적으로 고차원, 희소, 비선형적인 특성을 가진다. 이러한 특성에 따른 문제를 해결하기 위해 최종적으로 다양한 커널 함수를 사용할 수 있는 SVM 모델을 선정하여 모델을 만들었고, 비선형적 특성을 가지는 벡터를 결정 트리 기반으로 해결할 수 있으며 오타나 신조어를 규칙으로 파악이 가능한 XGBoost 모델을 선정하여 머신러닝을 진행하였다.

임베딩 기법과 모델을 선정해가는 과정과 최종 모델 또한 훈련 데이터의 개수는 욕설과 비욕설

각각 10000 개, 검증과 테스트 데이터는 그 10%인 1000 개를 사용하여 모델링을 진행하였다. 한글은 조사와 어미가 발달하여 이를 모두 고려하는 경우 '나/나의/나는/나를'를 모두 다른 토큰이 되기 때문에 모델의 정확성과 효율성에 영향을 미친다. 따라서 '은/는/이/가' 등의 조사와 분류에 필요하지 않은 '흠/음'과 같은 단어를 불용어 처리하였고 정규식을 활용하여 숫자 등 불필요한 단어를 제거한 뒤 한글과 영어만 남겨 모델링에 활용하였다.

비교를 위한 딥러닝 모델은 한국어 자연어 처리에 특화되지 않은 Bert 의 단점을 보완한 KoBert 모델을 선정하여 한국어 텍스트의 문맥을 통해 욕설을 분류해보고자 하였다. 욕설과 비욕설 훈련 데이터의 개수를 5000 개씩으로 줄였고 검증, 테스트 데이터는 각각 1000 개로 진행하였다.

## III. Result

최적의 모델을 찾기 위해 TF-IDF, fastText, doc2vec 세 가지의 임베딩 기법을 사용하여 각각 GridSearch 에 기반한 SVM 모델과 로지스틱 회귀 모델을 구현하였고 결과는 아래와 같다.

**표 1. TF-IDF 임베딩, SVM 모델 검증 평가 지표**

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.77 | 0.92 |
| Recall | 0.94 | 0.73 |
| F1-score | 0.85 | 0.81 |
| Accuracy | | 0.83 |

해당 모델의 테스트 정확도는 0.80 이었으며 욕설 분류에 있어 준수한 성능을 보였다.

**표 2. TF-IDF 임베딩 , 로지스틱 회귀 모델 검증 평가 지표**

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.77 | 0.92 |
| Recall | 0.94 | 0.72 |
| F1-score | 0.84 | 0.81 |
| Accuracy | | 0.83 |

해당 모델 또한 테스트 정확도에서 0.81 의 성능을 보이며 TF-IDF 임베딩이 욕설 분류에 있어 어느 정도의 성능 기대가 가능한 기법임을 알 수 있다.

**표 3. fastText 임베딩, SVM 모델 검증 평가 지표**

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.74 | 0.79 |
| Recall | 0.81 | 0.71 |
| F1-score | 0.77 | 0.75 |
| Accuracy | | 0.76 |

fastText 기법으로 임베딩 후 앞선 경우와 같은 SVM, 로지스틱 회귀 모델을 적용하였고 SVM 에서 나타난 검증 세트의 평가 지표는 [표 3]과 같았으며 해당 모델의 테스트 정확도는 0.76 으로 나타났다.

**표 4. fastText 임베딩, 로지스틱 회귀 모델 검증 평가 지표**

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.74 | 0.78 |
| Recall | 0.79 | 0.73 |
| F1-score | 0.76 | 0.75 |
| Accuracy | | 0.76 |

fastText 기법을 적용한 로지스틱 회귀 모델은 SVM 과 비슷한 검증 성능을 보였으며 테스트 정확도 또한 0.75 로 비슷하게 나타났다.

**표 5. doc2vec 임베딩, SVM 모델 검증 평가 지표**

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.68 | 0.82 |
| Recall | 0.87 | 0.60 |
| F1-score | 0.76 | 0.69 |
| Accuracy | | 0.73 |

doc2vec 기법을 적용한 SVM 모델에서는 앞선 두 임베딩 기법과 비교했을 때 욕설 분류에 좋지 못한 성능을 보였으며 특히 Recall 이 0.60 으로 실제 욕설을 정확하게 파악하지 못했다는 것을 알 수 있다. 또한 해당 모델의 테스트 정확도는 0.72 로 fastText 임베딩 모델에 비해 큰 차이를 보이지는 않았지만 검증 세트의 세부지표를 보았을 때 좋지 못한 방법임을 예상할 수 있었다.

**표 6. doc2vec 임베딩, 로지스틱 회귀 모델 검증 평가 지표**

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.68 | 0.72 |
| Recall | 0.74 | 0.66 |
| F1-score | 0.71 | 0.69 |
| Accuracy | | 0.70 |

[표 6]과 같이 로지스틱 회귀 모델에 적용하였을 때에도 세부 지표에서 앞선 결과들에 비해 좋지 못한 성능을 보였으며 테스트 정확도는 0.69 로 가장 낮은 성능을 보였다.

표 7. KoBert 딥러닝 평가 지표

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.89 | 0.91 |
| Recall | 0.91 | 0.89 |
| F1-score | 0.90 | 0.90 |
| Accuracy | | 0.90 |

앞선 머신러닝 모델들과 달리 딥러닝은 적은 데이터로 구현하였다. 절반 정도의 훈련 데이터를 활용하였음에도 불구하고 [표 7]과 같이 전반적으로 머신러닝 모델들보다 세부 지표에서 높은 성능을 보였으며 정확도 또한 0.90 으로 가장 높았다.

딥러닝을 제외한 6 개의 평가 지표를 바탕으로 TF-IDF 와 fastText 기법을 사용했을 때 상대적으로 좋은 성능을 보였음을 확인하였고, 두 기법을 병합하여 임베딩했을 때 저 좋은 성능에 도달할 수 있을 것이라 판단하여 최종 모델을 위한 기법으로 선정하였다.

해당 벡터의 특성을 바탕으로 로지스틱 회귀 대신 XGBoost 모델을 선정하였다. 최종 모델에서는 검증 세트의 평가 지표와 테스트 세트의 세부 평가 지표를 모두 도출하였으며 결과는 다음과 같다.

표 8. fastText+ TF-IDF 병합 임베딩, SVM 모델 검증 평가 지표

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.8114 | 0.8976 |
| Recall | 0.9087 | 0.7911 |
| F1-score | 0,8573 | 0.8410 |
| Accuracy | | 0.8496 |

표 9. fastText+ TF-IDF 병합 임베딩, SVM 모델 테스트 평가 지표

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.7939 | 0.8885 |
| Recall | 0.9037 | 0.7658 |
| F1-score | 0.8452 | 0.8226 |
| Accuracy | | 0.8347 |

[표 8], [표 9]를 보면 병합 임베딩을 거친 SVM 모델은 전반적으로 하나의 임베딩을 거친 모델보다 좋은 성능을 보였다. 이는 병합 임베딩이 성능 향상에 도움을 주었음을 보여준다. SVM 모델의 경우 테스트 세트의 정확도가 검증 세트의 정확도에 비해 조금 낮지만 그 차이가 크지 않으므로 과대적합이 일어났다고 보기는 어렵다. 테스트 세트의 Recall 이 다소 낮기는 하지만 전체적으로는 나쁘지 않은 성능을 보이므로 과소적합이 일어나지는 않았다.

표 10. fastText+ TF-IDF 병합 임베딩, XGBoost 모델 검증 평가 지표

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.7860 | 0.8741 |
| Recall | 0.8892 | 0.7606 |
| F1-score | 0.8345 | 0.8134 |
| Accuracy | | 0.8246 |

표 11. fastText+ TF-IDF 병합 임베딩, XGBoost 모델 테스트 평가 지표

| | 비욕설 | 욕설 |
|---|---|---|
| Precision | 0.7786 | 0.8897 |
| Recall | 0.9078 | 0.7423 |
| F1-score | 0.8382 | 0.8094 |
| Accuracy | | 0.8250 |

모델에 변화를 주어 비선형 벡터를 효과적으로 해결하도록 XGBoost 에 적용하였다. 결과는 [표 10], [표 11]과 같으며 이 또한 SVM 에 적용했을 때보단 낮지만 준수한 성능을 보인다는 점을 알 수 있었다. TF-IDF 만을 적용한 두 모델과 비교하였을 때 검증 세트에서는 성능이 비슷하거나 조금 낮지만 테스트 세트에서는 더 높은 정확도를 얻을 수 있었다. XGBoost 는 테스트 세트의 정확도가 검증세트에 비해 미세하게 높으므로 과대적합은 일어나지 않았다. 다만 XGBoost 또한 욕설이라 판단함에 있어 테스트 세트의 Recall 이 조금 낮은 것으로 보아 약간의 FN 이 발생할 수 있다는 가능성이 존재하지만, 충분히 높은 F1-score 과 정확도를 보이고 있으므로 과소적합이 일어나지 않았음을 확인할 수 있다.

## IV. Conclusion

본 프로젝트를 통해 다양한 모델과 임베딩 기법을 활용하여 텍스트 데이터에 포함된 욕설을 정확도 높게 분류하는 모델에 대해 알아보았다. 주어진 조건 하에 TF-IDF, fastText, doc2vec 의 임베딩을 시도했고, 이를 SVM, 로지스틱 회귀 모델에 적용하여 성능을 비교할 수 있었으며 이를 바탕으로 XGBoost 에도 적용하여 최종 모델을 만들어냈다. 비교를 위하여 딥러닝 또한 간소화하여 구현했으며 결과적으로 머신러닝은 약 83%, 딥러닝은 90%의 정확도를 얻을 수 있었다.

정확도에 있어서 크게 나쁘지 않은 결과를 얻을 수 있었지만, 프로젝트 과정과 결과에서 몇 가지 한계점을 찾을 수 있었고 향후 보완할 점을 알 수 있었다.

첫째, 데이터의 개수를 제한할 수밖에 없었다. 수집한 데이터는 약 500 만 개였다. 그러나 다양한 임베딩 기법을 활용하고, 다양한 모델을 활용하여 비교하여 최적의 머신러닝 모델을 찾는다는 본 연구의 목적에 따라 많은 경우의 수를 고려해야 하기 때문에

데이터의 수를 10000 개로 제한하여 모델을 구현하고 비교했다. 딥러닝보다 비용이 적은 머신러닝을 선택했더라도, 더 많은 수의 데이터를 활용하여 비교해보는 과정도 거쳤다면 더 좋은 성능을 찾아낼 수 있었을까 하는 의문을 남겼다. 이에 대해서는 더 효율적인 연구가 가능한 환경이 주어진다면 후속 연구로 해결할 수 있을 것이다.

둘째, 딥러닝이 적은 데이터일지라도 높은 성능을 보인다는 것을 직접 알아볼 수 있었다. 그러나 이를 통해 본 연구를 진행하기에는 시공간적 비용을 이겨내기 어려웠다. 5000 개의 데이터로 딥러닝 모델을 학습하는 데에도 적지 않은 시간이 걸렸기 때문에, 이를 비교용이 아닌 최종 모델로 결정했을 시에는 상당한 시간이 소요되었을 것으로 예상된다. 이러한 이유로 성능이 좋을 것으로 기대할 수 있는 부분에 대한 시도를 더욱 해보지 못한 것이다. 같은 주제를 가지고 딥러닝을 중심으로 한 연구가 필요함을 알 수 있었다.

셋째, 문맥과 상황을 고려하지 못한다는 것이다. 딥러닝 모델의 경우 문맥을 어느 정도 고려하지만 머신러닝 모델은 문맥이 아닌 단지 단어와 문장의 벡터를 가지고 학습하였으므로 욕설이 쓰인 상황을 반영하지 못한다. 욕설 또한 언어의 한 부분이므로 그 단어가 쓰이는 문맥과 상황은 분류에 굉장히 중요한 부분이다. 예를 들어, '시발점'이 욕설로 분류되는 것, 게임을 플레이하는 과정에서 뛰어난 실력을 보여준 유저에게 '미쳤다'라며 감탄하는 것은 욕설로 분류하기 모호한 부분이 존재한다. 물론 이 부분은 프로젝트에서 시행해보지는 않았다. 그러나 게임이나 문장의 의미를 이해하고 이를 반영하여 욕설을 분류할 수 있는 모델이 만들어진다면 상당히 수준 높고 기업들이 필요로 하는 모델이 될 수 있을 것이라 생각하였다.

본 프로젝트는 비용을 줄이기 위한 목적을 가지고 머신러닝을 통하여 여러 임베딩 기법과 모델로 많은 조합을 만들어 성능을 비교했고, 병합하여 사용함으로써 최적의 모델을 찾아갔다. 더하여 진행 과정 속 사회에 가장 필요한 모델을 위해서는 어떤 개선점이 필요한지 고민하고 발견했다는 점에서 의의를 가진다. 후속 연구에서 보다 개선된 환경을 바탕으로 한계점을 해결하고 높은 정확도를 도모할 필요가 있으며 온라인 공간 속 욕설과 비속어를 효과적으로 분류, 통제하는 모델을 만들 수 있을 것으로 기대한다.

## V. References


[1] 김규현, 서경희, 임시은. "인터넷 뉴스 댓글에서의 욕설의 분석." *The Sociolinguistic Journal of Korea*, 27(3), pp.63-96, 2019. http://dx.doi.org/10.14353/sjk.2019.27.3.03

[2] 이범준, 곽진아, 조성겸. "온라인 커뮤니티 이용자의 네트워크 지위와 욕설 사용의 관계: <일간베스트>를 중심으로." *사회과학연구*, 26(3), pp.391-416, 2015. http://dx.doi.org/10.16881/jss.2015.07.26.3.391

[3] 이주은. "욕설·비방 없는 게임 채팅창 만든다..넥슨의 AI 활용법" *한국금융신문*, 2023.09.07. https://www.fntimes.com/html/view.php?ud=202309072212552223959a82f9f5_18

[4] 송현섭. "KISO '욕설 필터링' 이용 37 개사 1 년간 욕설·비속어 33 만건 걸러내" *파이낸셜포스트*, 2024.08.29. https://www.financialpost.co.kr/news/articleView.html?idxno=212033

[5] 변희원. "미성년자에 19 금 채팅, 욕설까지…'성인용 AI'로 돈벌이 나선 빅테크" *조선일보*, 2025.04.28. https://www.chosun.com/economy/tech_it/2025/04/28/Y4QLXOAGAJCNPKCNBA3FQUCLWU/

[6] 이호석, 이홍래, 한요섭. "반자동 학습 기반의 비속어 및 욕설 탐지 시스템." *2017 년 한국컴퓨터종합학술대회 논문집*, pp.224-229, 2017.

[7] 박성희, 김휘강, 우지영. "딥러닝를 사용한 온라인 게임에서의 욕설 탐지." *한국컴퓨터정보학회 하계학술대회 논문집*, 27(2), pp.13-14, Jul. 2019.

[8] 이둠밈, 신영주. "HSV 스케일 이미지 처리를 이용한 텍스트 데이터의 한국어 욕설 탐지." *정보보호학회논문지 (Journal of The Korea Institute of Information Security & Cryptology)*, 35(2), pp.313-326, Apr. 2025. https://doi.org/10.13089/JKIISC.2025.35.2.313