온라인 채팅에서 그루밍 조기 감지

team 손은수 정해아

프로젝트 배경 및 목적

데이터 설명

모델 설명(Model Code)

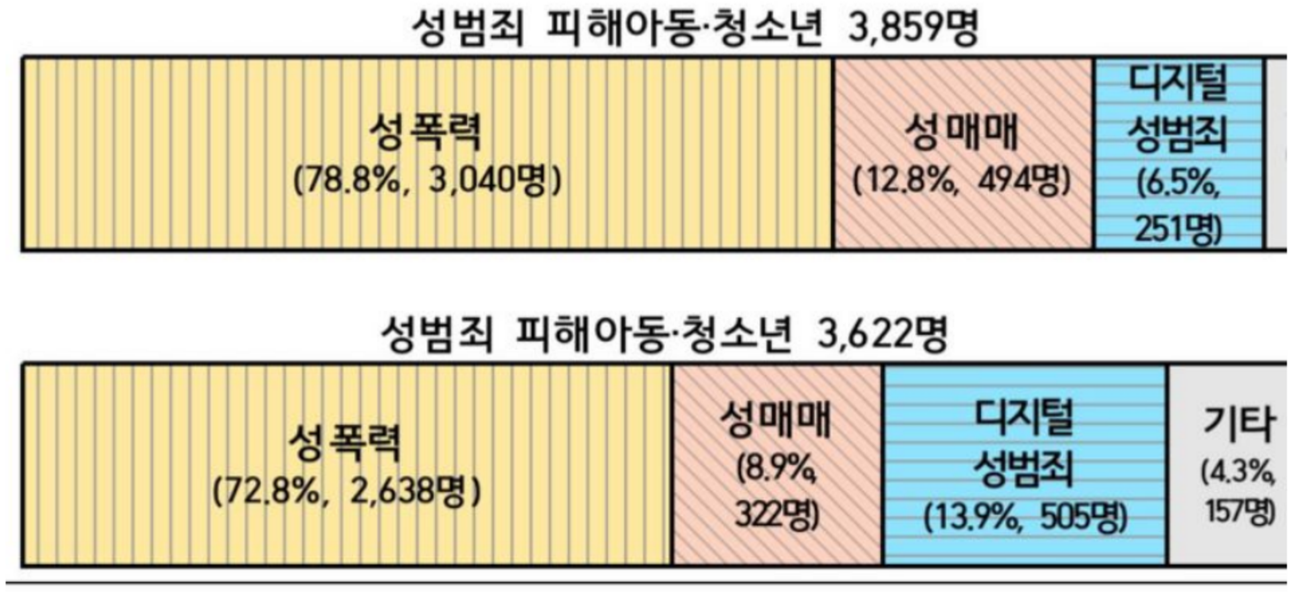
결과 설명 및 토의

Grooming

아동을 대상으로 친밀, 신뢰, 지배 관계를 설정해
성적 행동을 자연스럽게 받아들이게 하여 성적 접촉하는 행위

배경

< 2018~2019년 성범죄 유형별 피해아동·청소년 >



https://www.bbc.com/korean/news-56771525

만 14-18세 여성 청소년 온라인 그루밍 피해 현황

경험(대화내용)	경험율(%)
일상적인 대화나 관심사, 고민, 게임 관련 이야기를 함	97.4
나의 외모나 신체 대상으로 한 대화를 함	10.5
야한 농담이나 성적표현, 성행위 묘사 등의 대화를 함	7.9
나에게 본인 또는 제 3자의 야한 사진이나 음란물 전송 함	4.2
나의 얼굴 또는 전신이 담긴 일상에서 찍은 사진이나 동영상 요구 함	8.8
나에게 다리, 가슴, 성기 등의 신체 일부분의 노출이 담긴 사진이나 동영상 요구 함	4.0
내가 성적인 행위(자위행위 등)를 하도록 함	2.6
나에게 돈을 주거나 물건(게임아이템, 선물 등)을 사주며 성적인 대가 요구 함	1.6

※ 낯선 사람과 온라인 1:1 대화 경험율 : 70.0%

(자료: 여성가족부)

목적

-아동 성범죄를 예방

-사이버 그루밍과 관련된 행동 패턴을 식별

-텍스트 기반 대화를 분석하고 부적절한 언어나 개인 정보 요청과 같은 사이버 그루밍의 잠재적 지표를 식별

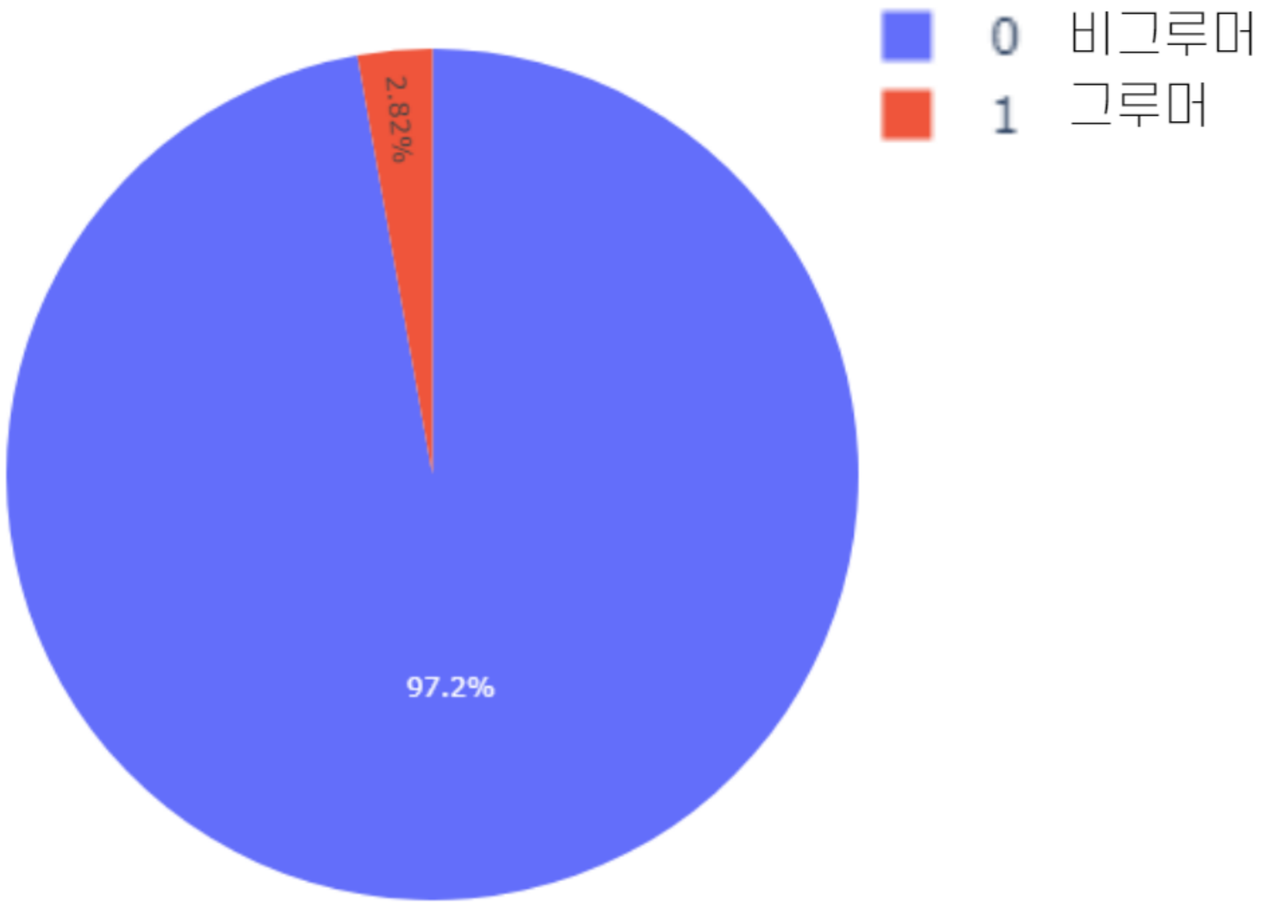
데이터 설명

- 오픈 소스 데이터 부족, 직접 수집 어려움
- 외국 데이터를 이용
- pan12대회에서 쓰여진 데이터와 2013년 이후의 데이터를 추가

- TP: (유죄) 그루밍 채팅 - 미성년 위장 피해자
- FP: (오탐) 성인간의 합의 성적대화
- TN: 주제 다양성을 위한 일반 채팅

```
<conversations>
  <conversation id="e621da5de598c9321a1d505ea95e6a2d">
    <message line="1">
      <author>97964e7a9e8eb9cf78f2e4d7b2ff34c7</author>
      <time>03:20</time>
      <text>Hola.</text>
    </message>
    <message line="2">
      <author>0158d0d6781fc4d493f243d4caa49747</author>
      <time>03:20</time>
      <text>hi.</text>
    </message>
```

http://www.perverted-justice.com/



-그루머 대화 3% 미만 -> 불균형 데이터

-현실 불균형성 반영

-채팅

불균형 데이터

불균형성 자체 해소
 -데이터 증식

불균형은 자연스러운 현실반영이다
 -이상탐지(Anomaly Detection) 모델
 -맞춤평가지표 -> f1 score

모델

1. Feature Extraction Model

2. Language Model

3. Ensemble

1. **Feature Extraction Model (BF-PSR)**

논문: How to take advantage of behavioral features for the early detection of grooming in online conversations

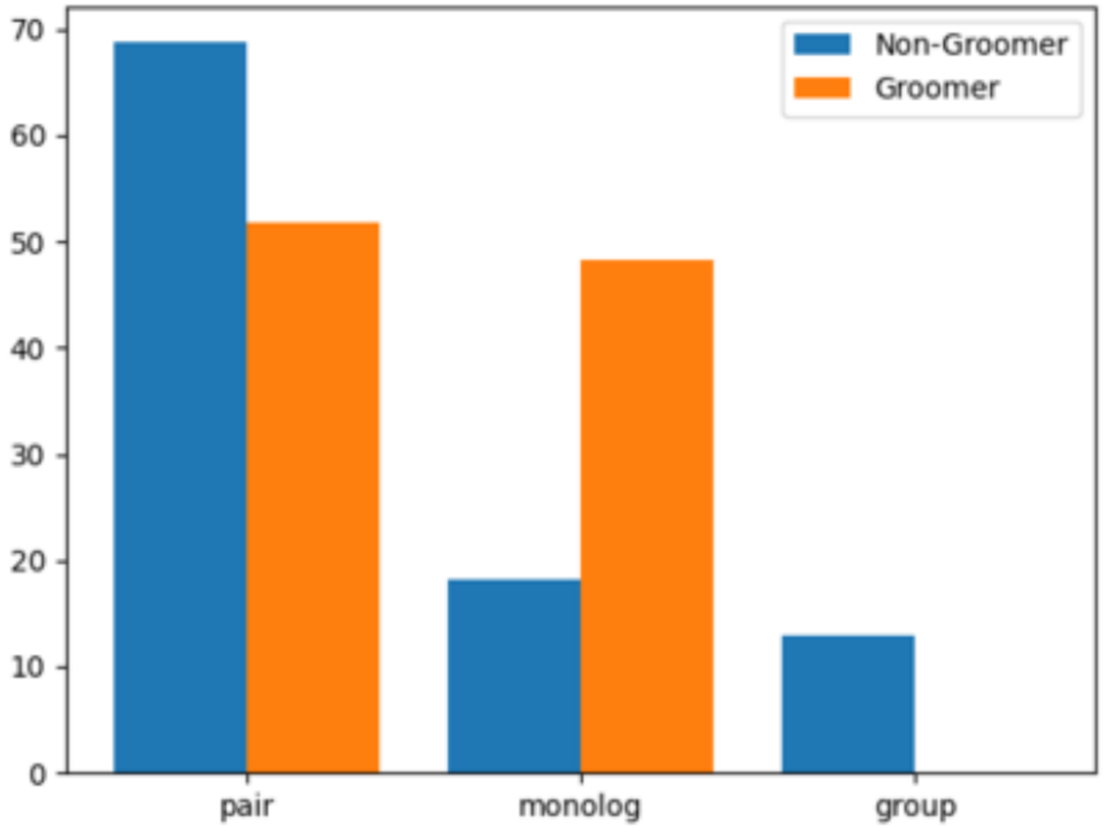
https://www.researchgate.net/publication/357415453_How_to_take_advantage_of_behavioral_features_for_the_early_detection_of_grooming_in_online_conversations

-non-sparse
-더 심도깊은 eda로 새로운 feature를 발견, 정의해 학습에 추가할 수 있다

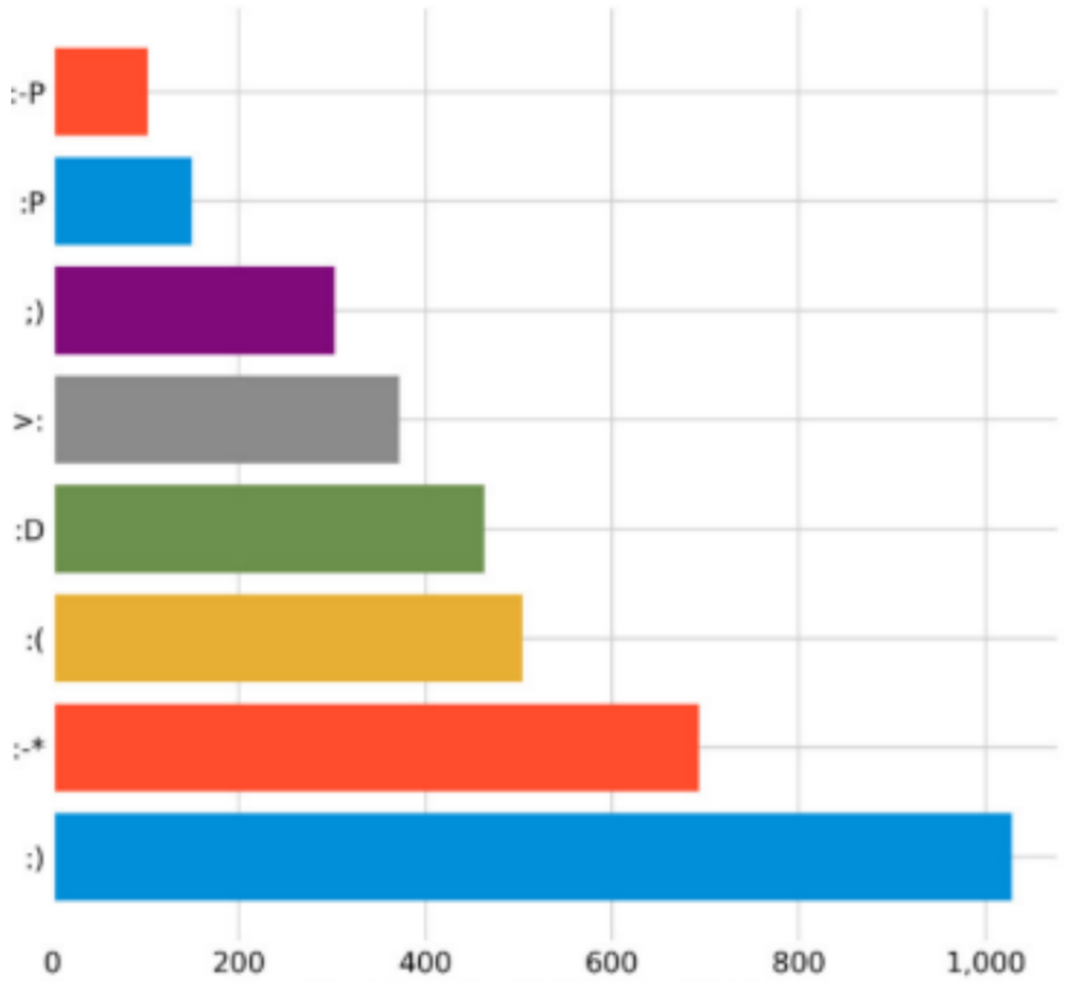
-데이터에서 끌어낼 수 있는 유용한 feature들을 모두 추출하지 못한다면 당연히 그렇게 소실된 정보만큼 bias가 발생한다

BF(Behavioral Features)

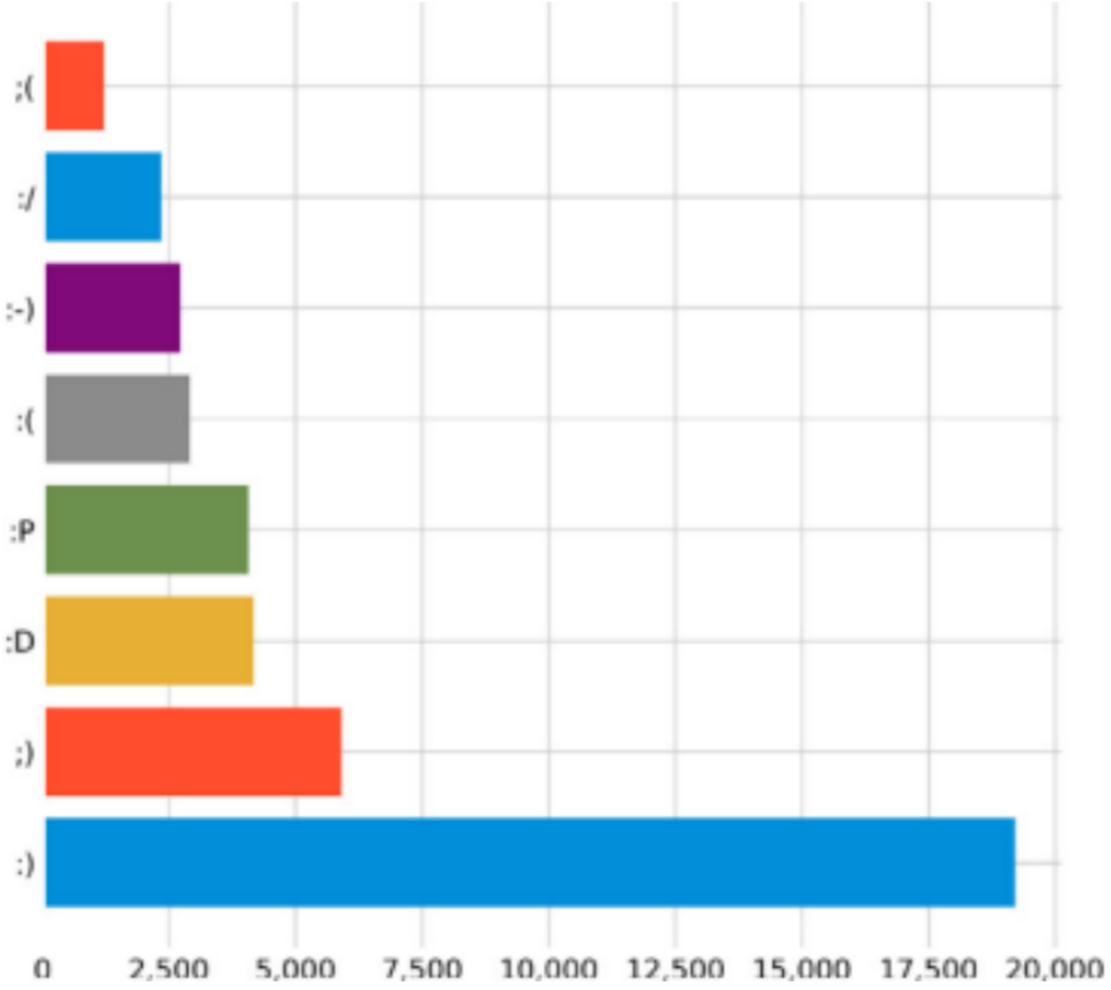
1. 대화 참여자 수



2. 이모티콘

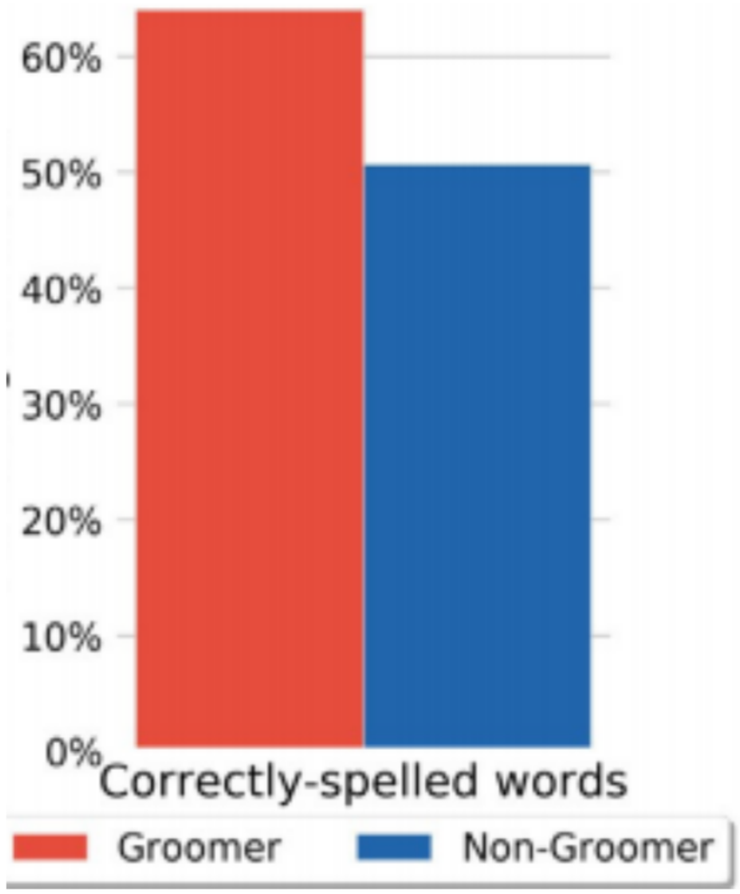


그루머



비그루머

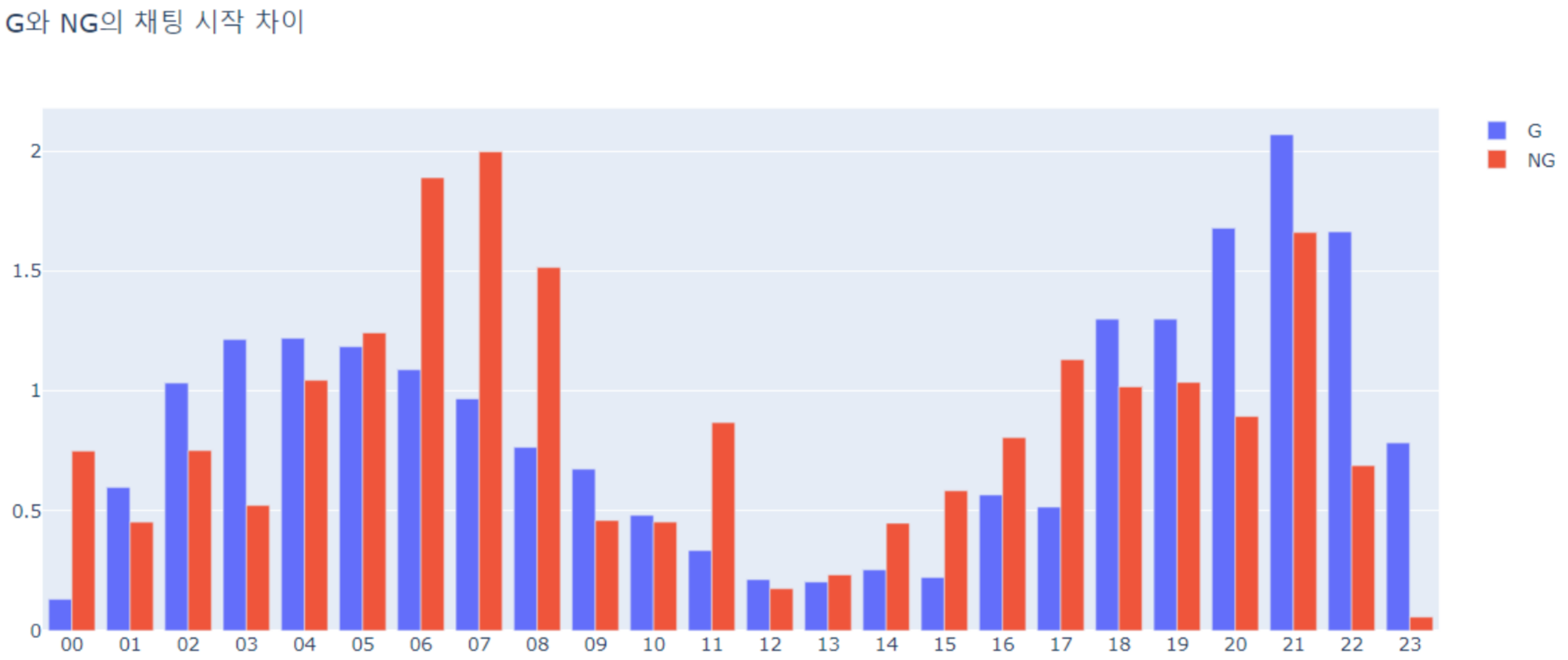
3. 맞춤법이 맞는 단어


60%
50%
40%
30%
20%
10%
0%
Correctly-spelled words
Groomer
Non-Groomer

4. 성적 주제 단어

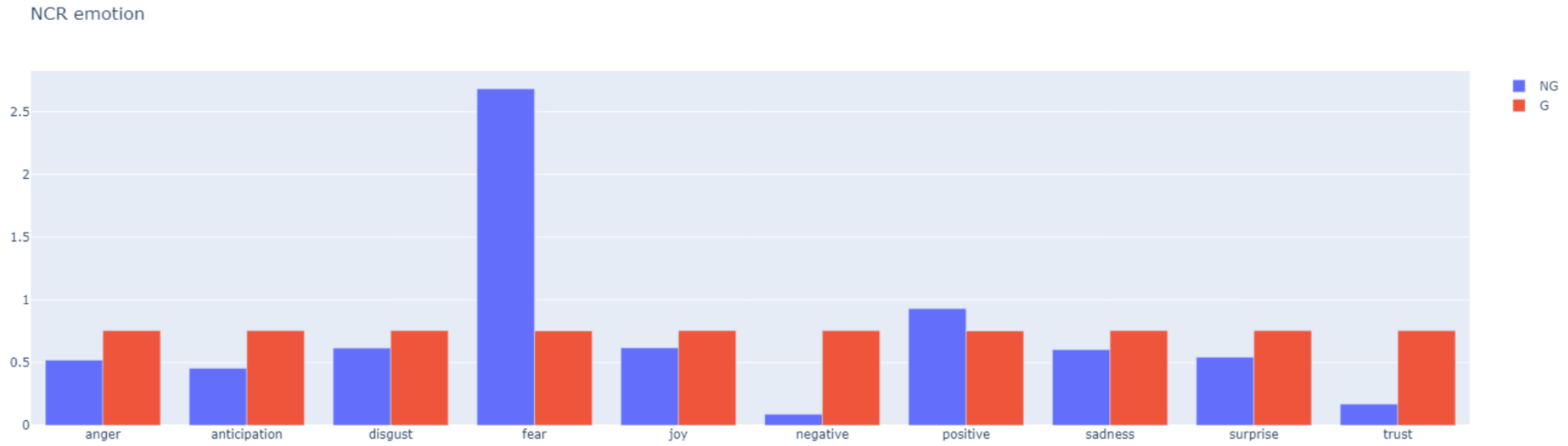

ur
sex
love
girl
baby
pussy
make
wan
kiss
sexy
fuck
naked
feel
meet
cock
hard
older
guy
hot
time
touch
horny
gon
EDIT
PORN
LINK
age
wish
nude
dick
suck
lot
lick
cool
stuff
young
kissing
wait
excited
panties
hope
show
bed
nice
play
big
pics
tongue
watch
trouble
talk

5. 대화가 시작되는 시간

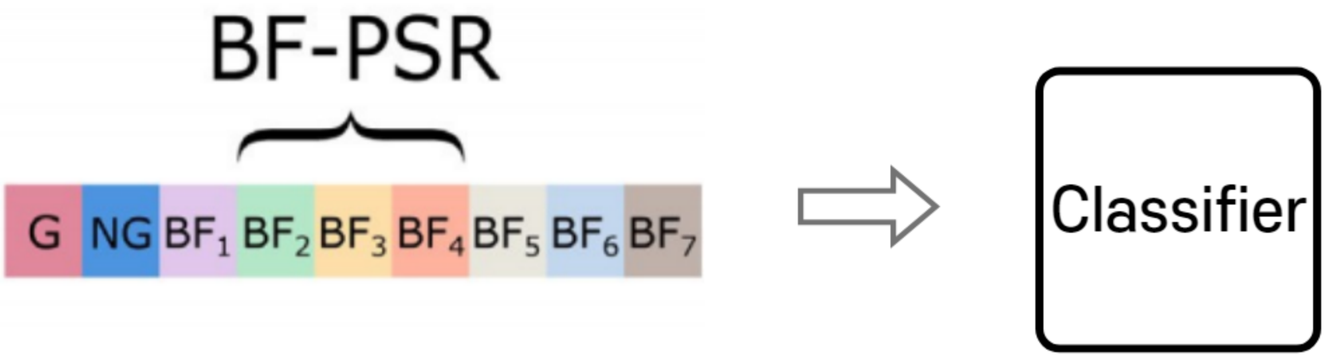

G와 NG의 채팅 시작 차이
G
NG
2
1.5
1
0.5
0
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23

6. 그루머와 비그루머의 성적 단어 차이

7. 감정 분석


NCR emotion
NG
G
2.5
2
1.5
1
0.5
0
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust

BF-PSR



TFIDF ->한 용어에 G와 NG 각각 가중치를 부여

가중치가 부여된 G와 NG를 7가지 행동 특징과 행렬로 연결해 BF-PSR 완성

MLP로 분류

F1 score 73%

2. Language Models

BERT vs GPT

Masked Language Model- 앞뒤문맥파악 Classification에 더 알맞음

BERT 계열중 어떠한 모델을 골라야하는가?

-비슷한 Task
-비슷한 Dataset
-비슷한 Input
-속도
etc

전처리? 모든 layer 학습? parameter? etc

2. Language Model

1. RoBERTa

-Bert 보다 ↑데이터 , 시간, batch-size, input sequence(512)
-동적 masking

-챗팅data 더 긴 input sequence 필요

2. bigbird-roberta-base

- sparse attention mechanism
-블록 내의 가까운 토큰간의 연결성 + 블록 간의 긴거리 연결성

3. longformer-base

-sliding window (local) attention + global attention
-long sequence- 4098 tokens

학습의 어려움

-너무 작은 data
-불균형 data
-FP 존재
-너무 큰 모델
-너무 긴 sequence (핵심 가해 문장/feature 대화안 비중 작을지도)
-비슷하지 않은 글 유형 (채팅) -vocab에 부재한 단어(unk token) 들 많을지도
-부족한 epoch 수

3. ensemble

Thank you

Thank you!