• Competition 주제 : jigsaw-unintended-bias-in-toxicity-classification

• 목표 : 온라인상의 악성 대화를 잡아내는 모델을 좀 더 정교하게 하여 낮은 에러율의 다양한 악성 대화를 잡아내는 모델을 만드는 것

• Data
- train.csv - 독성 레이블 및 하위 그룹을 포함하는 훈련 데이터
- test.csv - 독성 레이블 또는 하위 그룹을 포함 하지 않는 테스트 데이터
- sample_submission.csv - 샘플 제출 파일

※ 아래 데이터는 연구용으로 대회 종료후 추가된 데이터

- test_public_expanded.csv - 독성 라벨 및 하위 그룹을 포함한 공개 리더 보드 테스트 세트입니다. 경쟁 대상은> = 0.5 임계 값을 사용하여 쉽게 재구성 할 수있는 독성 컬럼의 이진화 된 버전이었습니다.
- test_private_expanded.csv - T독성 라벨 및 하위 그룹을 포함한 비공개 리더 보드 테스트 세트입니다. 경쟁 대상은> = 0.5 임계 값을 사용하여 쉽게 재구성 할 수있는 독성 컬럼의 이진화 된 버전이었습니다.
- toxicity_individual_annotations.csv - 독성 질문에 대한 개별 평가자 결정
1. id - 댓글 ID, train.csv, test_public_labeled.csv 또는 test_private_labeled.csv의 id 필드에 해당.
2. worker - 개별 주석자의 ID, 이러한 ID는 toxic_individual_annotations.csv 및 identity_individual_annotations.csv간에 공유됨.
3. toxic - 작업자가 댓글이 독성이라고 말하면 1, 그렇지 않으면 0.
4. severe_toxic - 작업자가 댓글이 심각하게 독성이라고 말하면 1, 그렇지 않으면 0. 심각한 독성으로 간주되는 모든 의견도 독성으로 간주됨.
5. identity_attack, insult, obscene, sexual_explicit, threat - 독성 하위 유형 속성. 작업자가 댓글이 이러한 각 특성을 나타내 었다고 말하면 1, 그렇지 않으면 0입니다.
- identity_individual_annoations.csv - 신원 질문에 대한 개별 평가자 결정.
1. id - 댓글 ID, train.csv, test_public_labeled.csv 또는 test_private_labeled.csv의 id 필드에 해당.
2. worker - 개별 주석자의 ID. 이러한 ID는 toxic_individual_annotations.csv 및 toxic_individual_annotations.csv간에 공유됨.

3. disability, gender, race_or_ethnicity, religion, sexual_orientation - 평가자가 댓글에서 발견 한이 범주 내의 정체성 목록. 공백으로 구분된 문자열을 형식화함.

- **중간 코드**

```
In [1]: import os
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import nltk
        nltk.download('stopwords')
        nltk.download('punkt')
        from nltk.corpus import stopwords
        from nltk.tokenize import word_tokenize
        stop_words = set(stopwords.words('english'))
        import warnings
        warnings.filterwarnings(action='ignore')
```

```
In [2]: os.chdir('C:/temp/Jigsaw Unintended Bias in Toxicity Classification')
```

```
In [3]: train_data = pd.read_csv('train.csv')
        test_data = pd.read_csv('test.csv')
        print(train_data.shape)
        print(test_data.shape)
```

```
(1804874, 45)
(97320, 2)
```

```
In [4]: train_data.head()
```

Out[4]:

	id	target	comment_text	severe_toxicity	obscene	identity_attack	insult	threat	asian	atheist	...	article_id	rating	funny	wow	sad	likes
0	59848	0.000000	This is so cool. It's like, 'would you want yo...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN	...	2006	rejected	0	0	0	0
1	59849	0.000000	Thank you!! This would make my life a lot less...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN	...	2006	rejected	0	0	0	0
2	59852	0.000000	This is such an urgent design problem; kudos t...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN	...	2006	rejected	0	0	0	0
3	59855	0.000000	Is this something I'll be able to install on m...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN	...	2006	rejected	0	0	0	0
4	59856	0.893617	haha you guys are a bunch of losers.	0.021277	0.0	0.021277	0.87234	0.0	0.0	0.0	...	2006	rejected	0	0	0	1

5 rows × 45 columns

In [5]:
```
train_df = train_data[['id','comment_text','target']]
test_df = test_data.copy()

# set index
train_df.set_index('id', inplace=True)
test_df.set_index('id', inplace=True)

# y_label
train_y_label = np.where(train_df['target'] >= 0.5, 1, 0)
train_df.drop(['target'], axis=1, inplace=True)
```

In [6]:
```
train_data['comment_text'].head(20)
```

Out[6]:
```
0     This is so cool. It's like, 'would you want yo...
1     Thank you!! This would make my life a lot less...
2     This is such an urgent design problem; kudos t...
3     Is this something I'll be able to install on m...
4                  haha you guys are a bunch of losers.
5                                ur a sh*tty comment.
6                          hahahahahahahahhha suck it.
7                                FFFFUUUUUUUUUUUUUUU
8     The ranchers seem motivated by mostly by greed...
9     It was a great show. Not a combo I'd of expect...
10                              Wow, that sounds great.
11    This is a great story. Man, I wonder if the pe...
12       This seems like a step in the right direction.
13    It's ridiculous that these guys are being call...
14    This story gets more ridiculous by the hour! A...
15    I agree; I don't want to grant them the legiti...
16    Interesting. I'll be curious to see how this w...
17                       Awesome! I love Civil Comments!
18    I'm glad you're working on this, and I look fo...
19    Angry trolls, misogynists and Racists", oh my....
Name: comment_text, dtype: object
```

In [7]:
```
# 문장 부호, stopwords 제거 class 정의
class clean_text:
    def __init__(self, text):
        self.text = text

    def rm_punct(text):
        punct = set([p for p in '/-'?!.,#$%₩'()*+-/:;<=>@[₩₩]^_`{|}~`'" + '""“”’‘´ ' + '∞θ÷α•à−β∅³ π ₹°£€₩×™√²—–&'])
        text = [t for t in text if t not in punct]
        return "".join(text)

    def rm_stopwords(text):
        word_tokens = word_tokenize(text)
        result = [w for w in word_tokens if w not in stop_words]
        return " ".join(result)
```

• 더 진행할 작업

- 데이터 토큰화 및 알맞은 분류 모델 만들기