

GoEmotions

- 58K Emotion Dataset from Reddit Comments -

GoEmotions: A Dataset of Fine-Grained Emotions

Dorottya Demszky^{1*} Dana Movshovitz-Attias² Jeongwoo Ko²

Alan Cowen² Gaurav Nemade² Sujith Ravi^{3*}

¹Stanford Linguistics ²Google Research ³Amazon Alexa

ddemszky@stanford.edu

{danama, jko, acowen, gnemade}@google.com

sravi@sravi.org

- Stanford Linguistics
- Google Research
 - Dataset 제작 지원

Abstract

Understanding emotion expressed in language has a wide range of applications, from building empathetic chatbots to detecting harmful online behavior. Advancement in this area can be improved using large-scale datasets with a fine-grained typology, adaptable to multiple downstream tasks. We introduce GoEmotions, the largest manually annotated dataset of 58k English Reddit comments, labeled for 27 emotion categories or Neutral. We demonstrate the high quality of the annotations via Principal Preserved Component Analysis. We conduct transfer learning experiments with existing emotion benchmarks to show that our dataset generalizes well to other domains and different emotion taxonomies. Our BERT-based model achieves an average F1-score of

Sample Text	Label(s)
OMG, yep!!! That is the final answer. Thank you so much!	gratitude, approval
I'm not even sure what it is, why do people hate it	confusion
Guilty of doing this tbph	remorse
This caught me off guard for real. I'm actually off my bed laughing	surprise, amusement
I tried to send this to a friend but [NAME] knocked it away.	disappointment

Table 1: Example annotations from our dataset.

sification into Ekman (Ekman, 1992b) or Plutchik (Plutchik, 1980) emotions.

TL;DR

- 58,000개의 Reddit English Comments
- 28개 Classes (= 27 emotions + Neutral)
- Multi-Label
- Bert 이용 시 F1 Score = 0.46

```
test.tsv x
data > test.tsv
1 I'm really sorry about your situation :( Although I love the names Sapphira, Cirilla, and Scarlett! 25 eecwqtt
2 It's wonderful because it's awful. At not with. 0 ed5f85d
3 Kings fan here, good luck to you guys! Will be an interesting game to watch! 13 een27c3
4 I didn't know that, thank you for teaching me something today! 15 eelgwd1
5 They got bored from haunting earth for thousands of years and ultimately moved on to the afterlife. 27 eem5uti
6 Thank you for asking questions and recognizing that there may be things that you don't know or understand about police tact
7 You're welcome 15 efdbh17
8 100%! Congrats on your job too! 15 ef0ec3b
9 I'm sorry to hear that friend :( It's for the best most likely if she didn't accept you for who you are 24 ee8utmi
10 Girlfriend weak as well, that jump was pathetic. 25 eeni74k
11 [NAME] has towed the line of the Dark Side. He wouldn't cross it by doing something like this. 3,10 eeaf7fg
12 Lol! But I love your last name though. XD 1,18 ee3yyz3
13 Translation }}} I wish I could afford it. 8 eegij2q
14 It's great that you're a recovering addict, that's cool. Have you ever tried DMT? 0,7 eeccgbb
15 I've also heard that intriguing but also kinda scary 14 edk4e66
16 I never wanted to punch osap harder after seeing that However not too hardly I cant afford them taking everything away 10
17 The thought of shooting anything at asylum seekers is appalling. 14 ed2e00i
18 if the pain doesn't go away after 4 hours or so, it's broke. 25,27 eezp1cd
19 Triggered:: Welp guess it's time for me to re-up lol 1 eearbor
```

1. Introduction

문제 제기

- Emotion Dataset의 한계
 - 데이터가 적고, 클래스도 많지가 않음
 - Ekman: joy, anger, fear, sadness, disgust, surprise (positive가 1개에 불과)
- Annotation의 Quality도 문제가 있음
- 한국어 역시 마찬가지 (KOSAC, NSMC)

개선점

- 레딧에서 58K의 Comment 가져옴 (데이터의 퀄리티를 고려하면 많은 양)
- 28개의 Emotion Class (Multi-Label)
- 기존 분류법(=taxonomy)은 positive가 1개인 것을 보완
- Data Analysis (with PPCA, Hierarchical Clustering)
- Bert-base-uncased로 F1 46%
- Transfer Learning

2. Related Works

2.1. Dataset

- 기존에는 뉴스 헤드라인이나 트위터를 이용
 - 트위터의 경우 emoji와 hashtag의 비중이 높음 (물론 reddit에도 없진 않다)
- Multiple annotation이 된 데이터가 없다

Ex) This makes me very happy, thank you for sharing! -> gratitude, joy

2.2. Taxonomy (분류법)

- 기존: Ekman(1992a)의 6가지 분류
 - joy, anger, fear, sadness, disgust, surprise
- Cowen이 제시한 방법을 text에 맞게 변형
 - 원 논문은 video, speech를 기준으로 taxonomy를 제시하여, 이 논문에서는 text에 대하여도 비슷하게 적용할 수 있다고 판단

2.3. Classification Model

- Feature-Based & Neural-Based
- 최근에는 BERT를 이용한 모델의 성능이 SOTA를 찍음
 - 감정 분류는 문맥이 중요하기 때문에 그럴 수 밖에 없을 듯

3. Dataset

3.1. 특징

- Reddit Comment (2005 ~ 2019)
 - <https://github.com/dewarim/reddit-data-tools>
- 27개 감정 + Neutral

Annotator에게 emoji도 같이 보여줌

admiration 🌟 Finding something impressive or worthy of respect.

amusement 😄 Finding something funny or being entertained.

anger 😡 A strong feeling of displeasure or antagonism.

annoyance 😞 Mild anger, irritation.

approval 👍 Having or expressing a favorable opinion.

caring Displaying kindness and concern for others.

confusion 😞 Lack of understanding, uncertainty.

curiosity A strong desire to know or learn something.

desire A strong feeling of wanting something or wishing for something to happen.

disappointment Sadness or displeasure caused by the nonfulfillment of one's hopes or expectations.

disapproval 🙅 Having or expressing an unfavorable opinion.

disgust 🤢 Revulsion or strong disapproval aroused by something unpleasant or offensive.

embarrassment 😳 Self-consciousness, shame, or awkwardness.

excitement 😄 Feeling of great enthusiasm and eagerness.

fear 😨 Being afraid or worried.

gratitude 🙏 A feeling of thankfulness and appreciation.

grief Intense sorrow, especially caused by someone's death.

joy 😄 A feeling of pleasure and happiness.

love ❤️ A strong positive emotion of regard and affection.

nervousness Apprehension, worry, anxiety.

optimism 🙌 Hopefulness and confidence about the future or the success of something.

pride Pleasure or satisfaction due to one's own achievements or the achievements of those with whom one is closely associated.

realization Becoming aware of something.

relief Reassurance and relaxation following release from anxiety or distress.

remorse Regret or guilty feeling.

sadness 😞 Emotional pain, sorrow.

surprise 😲 Feeling astonished, startled by something unexpected.

3.2. Preprocessing

- Reddit 데이터 자체가 Bias가 심함
 - young male user, offensive language
 - 내부적으로 가지고 있는 filter list를 이용
- 전처리
 - Manual review (욕설, 윤리적 이슈 방지)
 - Length filtering (NLTK, 3~30 tokens with median of 12)
 - Balancing
 - Masking ([NAME], [RELIGION])

3.3. Annotation

- 각 Example 당 3명의 rater가 평가
 - 인도 출신의 영어 Native Speaker
 - 미국 출신과 비교했을 때 emotion judgement가 비슷하다는 연구 결과 있음
- 우선 27개의 emotion의 정의를 보여주고, 예시문장도 함께 제공
- 여러 개의 emotion 선택 가능
- 어떠한 emotion인지 확신할 수 없을 때 Neutral
- 라벨을 달기 어렵다고 판단하면 no emotion이라고 체크 가능
 - 해당 example은 데이터셋에서 제거함

Number of examples	58,009
Number of emotions	27 + neutral
Number of unique raters	82
Number of raters / example	3 or 5
Marked unclear or difficult to label	1.6%
Number of labels per example	1: 83% 2: 15% 3: 2% 4+: .2%
Number of examples w/ 2+ raters agreeing on at least 1 label	54,263 (94%)
Number of examples w/ 3+ raters agreeing on at least 1 label	17,763 (31%)

- 82명의 rater 참여
- Unclear 비율 1.6%
- 대부분 emotion의 1개만 label

Table 2: Summary statistics of our labeled data.

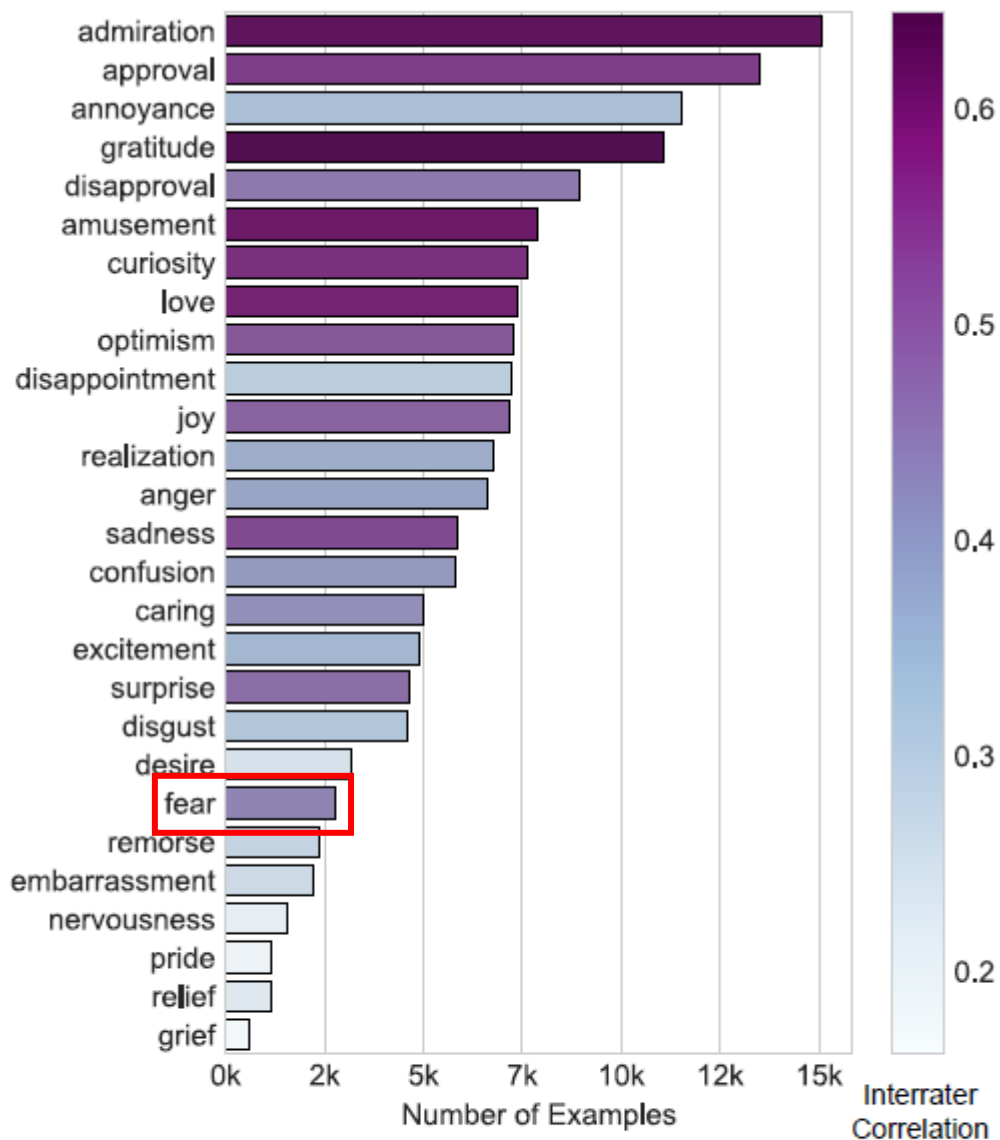
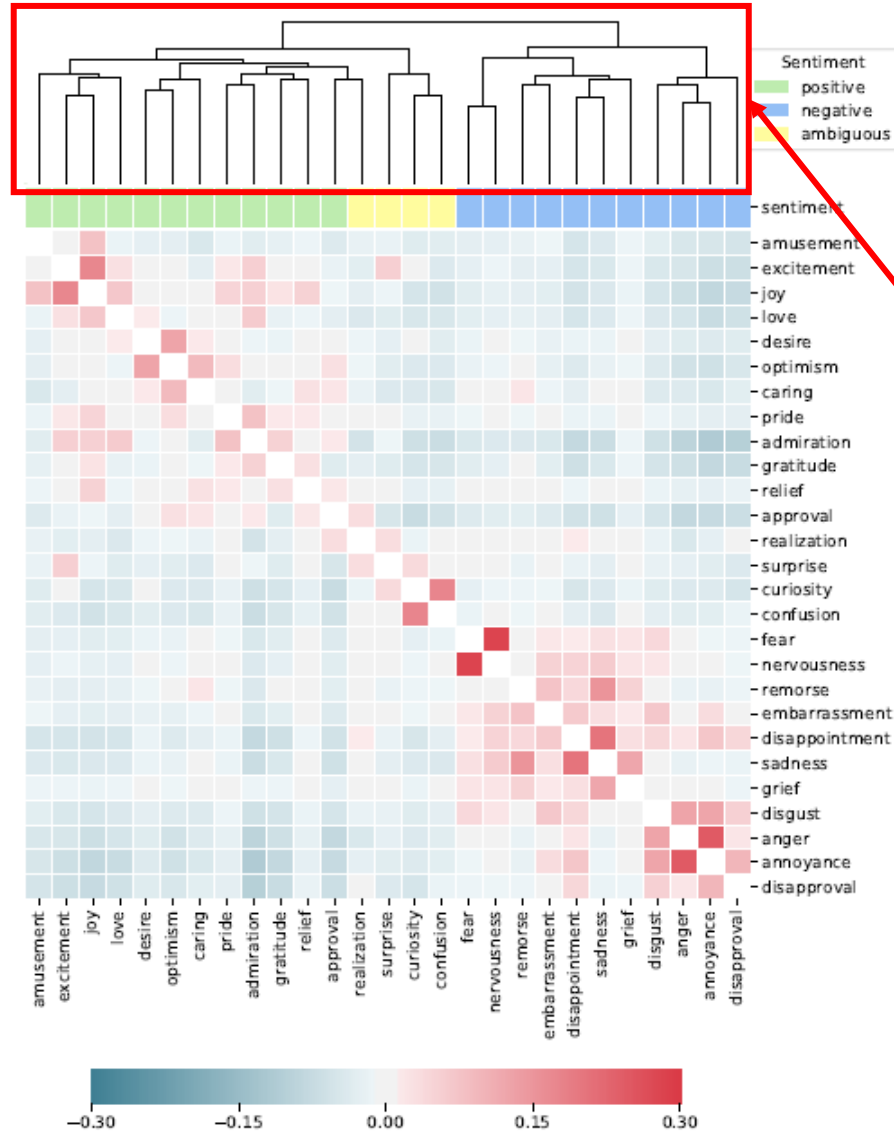


Figure 1: Our emotion categories, ordered by the number of examples where at least one rater uses a particular label. The color indicates the interrater correlation.

- Rating을 하기 전에 balancing을 했음에도 emotion frequency 차이가 존재
- Interrater Correlation
 - 각 emotion에 대해 rater들 간의 일치도(agreement)
 - Infrequent emotion에서도 interrater correlation이 높은 경우가 존재함
- [Personal Opinion]
 - 명확한 emotion(fear, sadness)에 대해서 이러한 경향이 보이는 듯. (= 이미 우리가 잘 알고 있는 감정)
 - 레딧 댓글 특성상 admiration이 많을 수 밖에 없을 듯 (페이스북에서도 좋은 글에 칭찬 남기듯이!!)



- Pearson Correlation을 이용하여 Emotion 간의 상관관계 분석
- Hierarchical Clustering

Figure 2: The heatmap shows the correlation between ratings for each emotion. The dendrogram represents the a hierarchical clustering of the ratings. The senti-ment labeling was done *a priori* and it shows that the clusters closely map onto sentiment groups.

4. Modeling

1. 28 emotion classes

admiration 🌟 Finding something impressive or worthy of respect.

amusement 😄 Finding something funny or being entertained.

anger 😡 A strong feeling of displeasure or antagonism.

annoyance 😞 Mild anger, irritation.

approval 👍 Having or expressing a favorable opinion.

caring Displaying kindness and concern for others.

confusion 😞 Lack of understanding, uncertainty.

curiosity A strong desire to know or learn something.

desire A strong feeling of wanting something or wishing for something to happen.

disappointment Sadness or displeasure caused by the nonfulfillment of one's hopes or expectations.

disapproval 🙅 Having or expressing an unfavorable opinion.

disgust 🤢 Revulsion or strong disapproval aroused by something unpleasant or offensive.

embarrassment 😳 Self-consciousness, shame, or awkwardness.

excitement 🥳 Feeling of great enthusiasm and eagerness.

fear 😨 Being afraid or worried.

gratitude 🙏 A feeling of thankfulness and appreciation.

grief Intense sorrow, especially caused by someone's death.

joy 😄 A feeling of pleasure and happiness.

love ❤️ A strong positive emotion of regard and affection.

nervousness Apprehension, worry, anxiety.

optimism 🙌 Hopefulness and confidence about the future or the success of something.

pride Pleasure or satisfaction due to one's own achievements or the achievements of those with whom one is closely associated.

realization Becoming aware of something.

relief Reassurance and relaxation following release from anxiety or distress.

remorse Regret or guilty feeling.

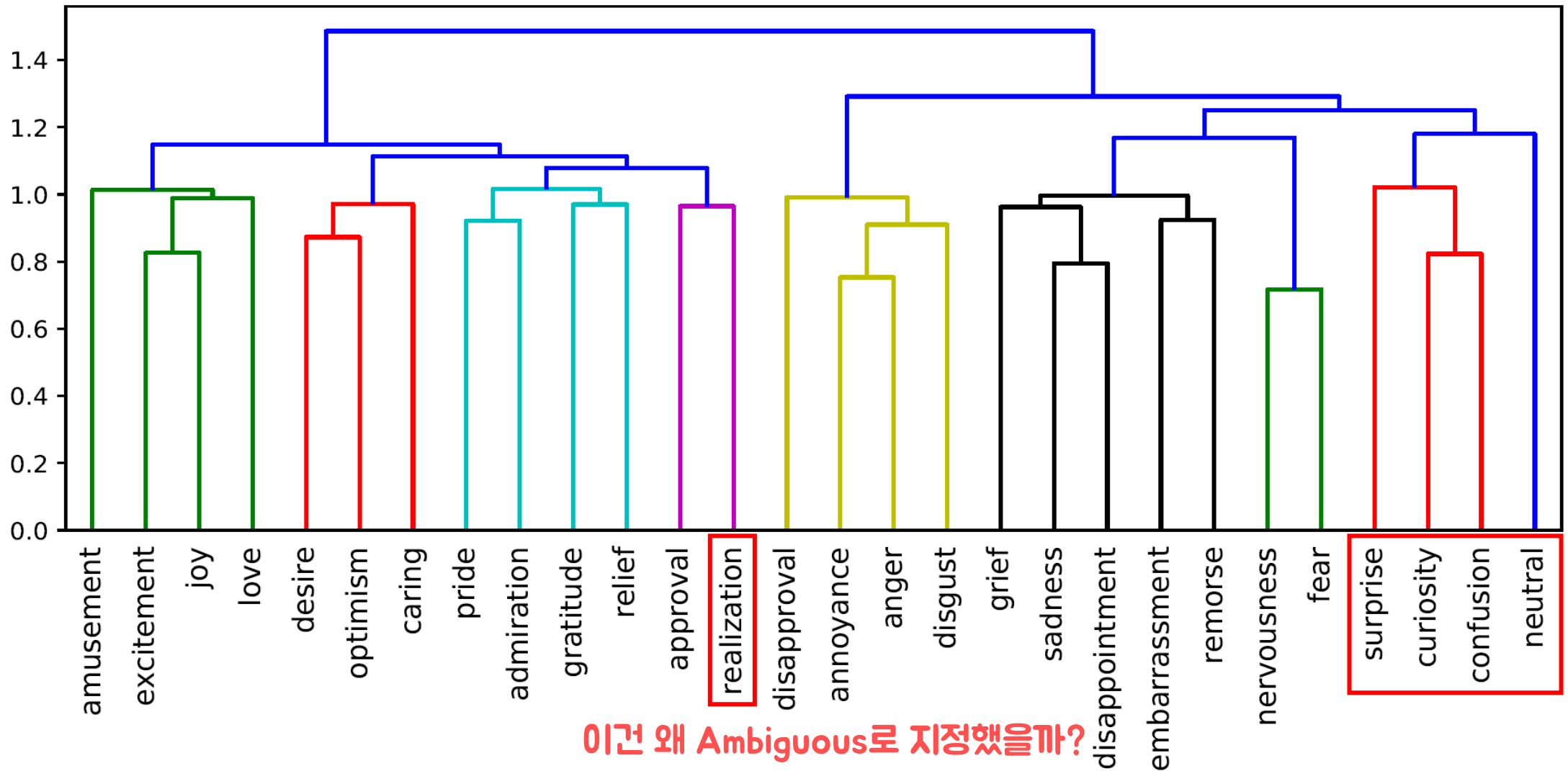
sadness 😞 Emotional pain, sorrow.

surprise 😲 Feeling astonished, startled by something unexpected.

2. Hierarchical Grouping

```
{
  "positive": [
    "amusement",
    "excitement",
    "joy",
    "love",
    "desire",
    "optimism",
    "caring",
    "pride",
    "admiration",
    "gratitude",
    "relief",
    "approval"
  ],
  "negative": [
    "fear",
    "nervousness",
    "remorse",
    "embarrassment",
    "disappointment",
    "sadness",
    "grief",
    "disgust",
    "anger",
    "annoyance",
    "disapproval"
  ],
  "ambiguous": ["realization", "surprise", "curiosity", "confusion"]
}
```

- Positive
 - Negative
 - Ambiguous
 - Neutral
-
- Ambiguous emotion도 대부분 positive sentiment text에서 많이 등장함



-> 논문에 ambiguous의 기준이 명확히 나와 있지 않았음

3. Ekman-Level (6 classes)

```
{  
  "anger": ["anger", "annoyance", "disapproval"],  
  "disgust": ["disgust"],  
  "fear": ["fear", "nervousness"],  
  "joy": [  
    "joy",  
    "amusement",  
    "approval",  
    "excitement",  
    "gratitude",  
    "love",  
    "optimism",  
    "relief",  
    "pride",  
    "admiration",  
    "desire",  
    "caring"  
  ],  
  "sadness": ["sadness", "disappointment", "embarrassment", "grief", "remorse"],  
  "surprise": ["surprise", "realization", "confusion", "curiosity"]  
}
```

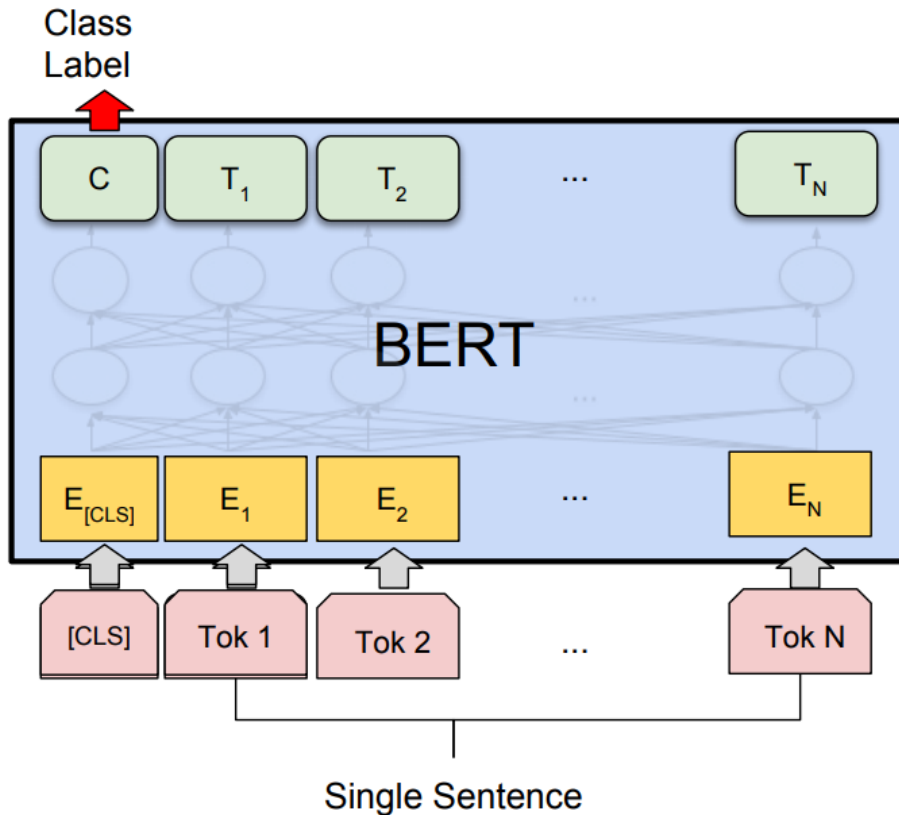
Ambiguous를 다 여기로 보냄....

Model Baseline



Google
BERT

Model Baseline



- 일반적으로 여러 개의 class 중에서 1개 만을 예측하는 것이 아닌 여러 개의 class를 예측하는 것
- Softmax가 아닌 Sigmoid를 적용 (threshold=0.3)
- <https://github.com/monologg/GoEmotions-Korean/blob/master/model.py>

Emotion	Precision	Recall	F1
admiration	0.53	0.83	0.65
amusement	0.70	0.94	0.80
anger	0.36	0.66	0.47
annoyance	0.24	0.63	0.34
approval	0.26	0.57	0.36
caring	0.30	0.56	0.39
confusion	0.24	0.76	0.37
curiosity	0.40	0.84	0.54
desire	0.43	0.59	0.49
disappointment	0.19	0.52	0.28
disapproval	0.29	0.61	0.39
disgust	0.34	0.66	0.45
embarrassment	0.39	0.49	0.43
excitement	0.26	0.52	0.34
fear	0.46	0.85	0.60
gratitude	0.79	0.95	0.86
grief	0.00	0.00	0.00
joy	0.39	0.73	0.51
love	0.68	0.92	0.78
nervousness	0.28	0.48	0.35
neutral	0.56	0.84	0.68
optimism	0.41	0.69	0.51
pride	0.67	0.25	0.36
realization	0.16	0.29	0.21
relief	0.50	0.09	0.15
remorse	0.53	0.88	0.66
sadness	0.38	0.71	0.49
surprise	0.40	0.66	0.50
macro-average	0.40	0.63	0.46
std	0.18	0.24	0.19

Table 4: Results based on GoEmotions taxonomy.

Sentiment	Precision	Recall	F1
ambiguous	0.54	0.66	0.60
negative	0.65	0.76	0.70
neutral	0.64	0.69	0.67
positive	0.78	0.87	0.82
macro-average	0.65	0.74	0.69
std	0.09	0.10	0.09

Table 5: Results based on sentiment-grouped data.

Ekman Emotion	Precision	Recall	F1
anger	0.50	0.65	0.57
disgust	0.52	0.53	0.53
fear	0.61	0.76	0.68
joy	0.77	0.88	0.82
neutral	0.66	0.67	0.66
sadness	0.56	0.62	0.59
surprise	0.53	0.70	0.61
macro-average	0.59	0.69	0.64
std	0.10	0.11	0.10

Table 6: Results using Ekman’s taxonomy.

예제 코드 (영어 원본 데이터)

<https://github.com/monologg/GoEmotions-pytorch>

- 논문에서 언급된 3가지 분류법으로 모두 시도함
 - Original Goemotions, Hierarchical Grouping, Ekman
 - Huggingface s3를 통해 Pipeline 제공

예제 코드 (한국어 번역 데이터)

<https://github.com/monologg/GoEmotions-Korean>

- 구글 번역기로 데이터를 한국어로 번역
 - 정확도는 구글 번역기, 자연스러움은 파파고
 - 레딧 댓글인 관계로 번역 품질이 매우 매우 안 좋음...
- Transformers 라이브러리를 이용하여 제작
 - BERT 이후의 최신 아키텍처에도 테스트 가능

📄 텍스트

📄 문서

영어 - 감지됨

영어

한국어

독일어

▼

↔

한국어

영어

일본어



▼


Man I love Canada🇨🇦


×




내가 캐나대를 사랑하는 사람

☆

21/5000 



```
from multilabel_pipeline import MultiLabelPipeline
from transformers import ElectraTokenizer
from model import ElectraForMultiLabelClassification
from pprint import pprint

tokenizer = ElectraTokenizer.from_pretrained("monologg/koelectra-base-finetuned-goemotions")
tokenizer.add_special_tokens({"additional_special_tokens": ["[NAME]", "[RELIGION]"]}) # BUG: It should
be hard-coded on transformers v2.9.1
model = ElectraForMultiLabelClassification.from_pretrained("monologg/koelectra-base-finetuned-
goemotions")

goemotions = MultiLabelPipeline(
    model=model,
    tokenizer=tokenizer,
    threshold=0.3
)

texts = [
    "전혀 재미 있지 않습니다 ...",
    "나는 “지금 가장 큰 두려움은 내 상자 안에 사는 것” 이라고 말했다.",
    "곱창... 한시간반 기다릴 맛은 아님!",
    "애정하는 공간을 애정하는 사람들로 채울때",
    "너무 좋아",
    "답러닝을 짝사랑중인 학생입니다!",
    "마음이 급해진다.",
    "아니 진짜 다들 미쳤나봐ㅋㅋㅋ",
    "개노잼"
]

pprint(goemotions(texts))

# Output
[{'labels': ['disapproval'], 'scores': [0.82489157]},
 {'labels': ['fear'], 'scores': [0.9509703]},
 {'labels': ['neutral'], 'scores': [0.9585297]},
 {'labels': ['approval', 'neutral'], 'scores': [0.62351847, 0.34225133]},
 {'labels': ['admiration'], 'scores': [0.97146636]},
 {'labels': ['love', 'neutral'], 'scores': [0.32616842, 0.5455638]},
 {'labels': ['caring', 'nervousness'], 'scores': [0.51289016, 0.4741806]},
 {'labels': ['amusement'], 'scores': [0.9680228]},
 {'labels': ['anger', 'annoyance'], 'scores': [0.5345557, 0.764603]}]
```

감사합니다!

2020.07.09 박장원