ZeroToOne

Start

DLThon

박혜원, Gyubin Hwang, Yeonsoo Kim, Bumjun Kim

Task: TEXT CLASSIFICATION

OVERVIEW

01
OBJECTIVES
Lorem ipsum dolor sit

02
PROBLEM
Lorem ipsum dolor sit

03
LITERARY PREVIEW
Lorem ipsum dolor sit

04
THEORETICAL
Lorem ipsum dolor sit

05
OBJECTIVES
Lorem ipsum dolor sit

06
HYPOTHESIS
Lorem ipsum dolor sit

07
METHODOLOGY
Lorem ipsum dolor sit

08
IMPLEMENTATION
Lorem ipsum dolor sit

09
RESULT
Lorem ipsum dolor sit

10
CONCLUSION
Lorem ipsum dolor sit

11
RECOMMENDATIONS
Lorem ipsum dolor sit

BUSINESS MARKETING

APPROACH

평가 기준

1. 데이터 EDA와 데이터 전처리가 적절하게 이뤄졌는가?
2. task에 알맞게 적잘한 모델을 찾아보고 선정했는가?
3. 성능향상을 위해 논리적으로 접근했는가?
4. 결과 도출을 위해 여러 가지 시도를 진행했는가?
5. 도출된 결론에 충분한 설득력이 있는가?
6. 적절한 metric을 설정하고 그 사용 근거 및 결과를 분석하였는가?
7. 발표가 매끄럽게 진행되었고 발표시간을 준수하였는가?

KEYWORDS

EDA (Exploratory Data Analysis)
Preprocessing
Model
Performance
Metrics

INITIAL ACTION

Directly Exploring and Analyzing Data for EDA

DATA

Directly Exploring and Analyzing Data for EDA

```
2,기타 괴롭힘 대화,"너 되게 귀여운거 알지? 나보다 작은 남자는 첨봤어.
그만해. 니들 놀리는거 재미없어.
지영아 너가 키 160이지? 그럼 재는 160도 안돼는거네?
너 군대도 안가고 좋겠다.
니들이 나 작은데 보태준거 있냐?
난쟁이들도 장가가고하던데. 너도 희망을 가져봐
더이상 하지마라.
그 키크는 수술도 있대잖아? 니네 엄마는 그거 안해주디?
나람 해줬어. 저 키로 어찌살아.
제발 그만 괴롭히라고!"
3,갈취 대화,"어이 거기
예??
너 말이야 너. 이리 오라고
무슨 일.
너 옷 좋아보인다?
얘 돈 좀 있나봐
아니에요.돈 없어요
뒤져서 나오면 넌 죽는다
오늘 피시방 콜?
콜. 마지막 기회다. 있는거 다 내놔
정말 없어요"
```

t_001/text	2	우리팀에서 다른팀으로 갈 사람 없나? 그럼 영지씨가 가는건 어때? 네? 제가요? 그렇지? 2달만 파견 잘 갔다오면 승진이야. 네? 저는 별로 가고 싶지 않습니다. 여기 있는 모든사람도 가기 싫어해. 그
t_002/text	3	너 오늘 그게 뭐야 네 제가 뭘 잘못했나요.? 제대로 좀 하지 네 똑바로 좀 하지 행실이 맘에 안들어 네 실천할께요 옆에 동기들 좀 본받으란 말야 어? 네 알겠습니다 그래 똑바로해 노력할께요
t_003/text	4	이거 들어바 와 이 노래 진짜 좋다 그치 요즘 이 것만 들어 진짜 너무 좋다 내가 요즘 듣는 것도 들어봐 음 난 좀 별론데 좋을 줄 알았는데 아쉽네 내 취향은 아닌 듯 배고프다 밥이나 먹으러 가자 그래
t_004/text	3	아무튼 앞으로 니가 내 와이파이야. .응 와이파이 온. 켰어. 반말? 주인님이라고도 말해야지? .켰습니다. 주인님. 오냐. 앞으로도 잘부탁한다 와이파이. 넵. 이렇게만 해? 알겠지? 응.
t_005/text	0	그러니까 빨리 말해. 선생님 제발 살려주십시오. 비밀번호 틀릴 때마다 손톱 하나씩 뺀찌로 뽑는다. 선생님 저도 정말 모릅니다. 하나 엄지 손톱 뽑는다. 으악! 잘못했습니다. 그런데 진짜 모릅니다.
t_006/text	0	그 사람 누구냐? 누구말하는거야? 다 알면서 모른척 하지마. 둘다 쏴버리기 전에. 아니 정말 왜이래? 누굴말하는거냐고. 어제 밤에 전화온 그새끼 말이야. 번호 알아놨으니까 죽이기 전에 다 불어라.
t_007/text	1	야 저기야 닭꼬치 트럭왔다 응 그러네 그치? 너도 먹고 싶지? 어?나도? 그래 너 세상에 닭꼬치 싫어하는 사람이 어딨어 그렇긴한데 배부른데 그래? 그럼 나만 먹고 올게 응. 뭐해?? 뭘? 돈줘야 먹고오

MACHINE LEARNING – PREPROCESSING

```
train_data = train_data.sample(frac = 1, random_state = 42)

def preprocess_sentence(sentence):

    sentence = sentence.lower().strip()
    sentence = re.sub(r"([?.!,])", r" \1 ", sentence)
    sentence = re.sub(r'[" "]+', " ", sentence)
    sentence = re.sub(r"[^a-zA-Z?.!가-힣ㄱ-ㅎㅏ-ㅣ]+", " ", sentence)
    sentence = sentence.strip()

    return sentence
```

1. Transform lowercase letters, remove spaces

2. . ? ! Handles spaces before and after punctuation marks such as , etc.

3. If there are two or more spaces, one space is processed.

4. Remove characters other than a~z, A~Z, ?, ., !, 가~힣, ㄱ~ㅎ, ㅏ~ㅣ, etc.

5. remove spaces

MACHINE LEARNING – PREPROCESSING

```
def check_class(it):
    if '협박' in it:
        return 0
    elif '갈취' in it:
        return 1
    elif '직장 내 괴롭힘' in it:
        return 2
    elif '기타 괴롭힘' in it:
        return 3
```

The four classes, including intimidation, extortion, workplace harassment, and other harassment, were assigned 0, 1, 2, and 3, respectively.

COMPARISON

ML VS KLUE/BERT-BASE

```
Linear Support Vector Machine Accuracy: 0.8291139240506329
Logistic Regression Accuracy: 0.8063291139240506
Decision Trees Accuracy: 0.6354430379746835
Random Forest Accuracy: 0.7556962025316456
K-Nearest Neighbors Accuracy: 0.7240506329113924
Naive Bayes Accuracy: 0.8354430379746836
Gradient Boosting Accuracy: 0.7506329113924051
Linear Discriminant Analysis Accuracy: 0.46455696202531643
```

LLM Pretrained Models	klue-bert 5 epoch	klue-bert 3 epoch	klue-bert 4 epoch
Data Original	val_loss: 0.4654 - val_accuracy: 0.8962 ACCURACY : 0.9	val_loss: 0.4078 - val_accuracy: 0.8886 ACCURACY : 0.9025	al_loss: 0.4928 - val_accuracy: 0.8797 ACCURACY: 0.895

The Naive Bayes model, exhibiting the highest accuracy, achieves an approximate accuracy rate of 83.5%, whereas the Klue/bert-base model attains a higher accuracy at around 90%.

PRE-TRAINED MODELS

Baseline model: Klue/bert-base

Trial:

- Klue/roberta-large
- beomi/KcELECTRA-small

REASON FOR SELECTION

https

Thynk University

Start

PRE-TRAINED MODELS

KLUE/BERT-BASE

Subject : BUSINESS MARKETING

Submit by : CLAUDIA ALVES

Thynk University

Start

PRE-TRAINED MODELS

SKT/KOGPT2-BASE-V2E

제가 사용한 실제 이유
1. 학습이 빨라서,,
2. Enc + Dec 구조인 모델은 성능이 어떤지 확인하고자

추가 조사 설명 based on : https://github.com/SKT-AI/KoGPT2

Subject : BUSINESS MARKETING | TEST ACCURACY: 0.88

Thynk University

Start

PRE-TRAINED MODELS

KCELECTRO-SMALL

Presentation are communication tools that can be used as demontrations, lectures, reports, and more. it is mostly presented before an audience.

Subject : BUSINESS MARKETING

Submit by : CLAUDIA ALVES

Thynk University

Start

PRE-TRAINED MODELS

MONOLOGG/KOELECTRA BASE-V3

Presentation are communication tools that can be used as demontrations, lectures, reports, and more. it is mostly presented before an audience.

Subject : BUSINESS MARKETING

Submit by : CLAUDIA ALVES

INSUFFICIENT DATA

클래스	Class No.	# Training	# Test
협박	00	896	100
갈취	01	981	100
직장 내 괴롭힘	02	979	100
기타 괴롭힘	03	1,094	100

DATA AUGMENTATION

KorEDA

AI Hub Data

Back Translation

KorEDA

이 프로젝트는 EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks 를 한국어로 쓸 수 있도록 wordnet 부분만 교체한 프로젝트 입니다.

wordnet은 KAIST에서 만든 Korean WordNet(KWN) 을 사용했습니다.

github.com/catSirub/KorEDA

```python
def EDA(sentence, alpha_sr=0.1, alpha_ri=0.1, alpha_rs=0.1, p_rd=0.1, num_aug=9):
        sentence = get_only_hangul(sentence)
        words = sentence.split(' ')
        words = [word for word in words if word is not ""]
        num_words = len(words)

        augmented_sentences = []
        num_new_per_technique = int(num_aug/4) + 1

        n_sr = max(1, int(alpha_sr*num_words))
        n_ri = max(1, int(alpha_ri*num_words))
        n_rs = max(1, int(alpha_rs*num_words))

        # sr
        for _ in range(num_new_per_technique):
                a_words = synonym_replacement(words, n_sr)
                augmented_sentences.append(' '.join(a_words))

        # ri
        for _ in range(num_new_per_technique):
                a_words = random_insertion(words, n_ri)
                augmented_sentences.append(' '.join(a_words))

        # rs
        for _ in range(num_new_per_technique):
                a_words = random_swap(words, n_rs)
                augmented_sentences.append(" ".join(a_words))

        # rd
        for _ in range(num_new_per_technique):
                a_words = random_deletion(words, p_rd)
                augmented_sentences.append(" ".join(a_words))
```

KorEDA – Hyperparameters

sr **synonym replacement**	**ri** **random insertion**
rs **random swap**	**rd** **random deletion**

https://github.com/jasonwei20/eda_nlp

- **Synonym Replacement (SR):** Randomly choose *n* words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random.
- **Random Insertion (RI):** Find a random synonym of a random word in the sentence that is not a stop word. Insert that synonym into a random position in the sentence. Do this *n* times.
- **Random Swap (RS):** Randomly choose two words in the sentence and swap their positions. Do this *n* times.
- **Random Deletion (RD):** For each word in the sentence, randomly remove it with probability *p*.

sr
synonym replacement

ri
random insertion

rs
random swap

rd
random deletion

원문 데이터

제가 우울감을 느낀지는 오래됐는데 점점 개선되고 있다고 느껴요

data augmentation한 데이터

우울감을 느낀지는 오래됐는데 점점 개선되고 있다고
제가 우울감을 느낀지는 오래됐는데 느껴요 개선되고 있다고 점점
오래됐는데 우울감을 느낀지는 제가 점점 개선되고 있다고 느껴요
느껴요 우울감을 느낀지는 오래됐는데 점점 개선되고 있다고 제가

Example 1: 씨 이번에 또 틀렸잖아 . 죄송합니다 . 다시 확인하겠습니다 . 동기인 씨는 한번 알려주면 끝인데 자네는 매번 알려줘도 이래 . 죄송하지만 팀장님한테 확인받고 한 작업입니다 . 뭐 그럼 내가 일부러 그런다는 거야 . 그건 아니지만 . 그런 식으로 일할거면 회 관둬요 . 누가 보면 내가 괴롭힌다고 생각하겠어 그런 말씀은 지나치십니다 . 그럼 알아서 잘 하던가 . 눈치도 없이 . 다음부터는 다시 확인하겠습니다 . 팀장님도 말씀 가려서 해주세요 .
Example 2: 씨 이번에 또 틀렸잖아 . 죄송합니다 . 다시 확인하겠습니다 . 동기인 씨는 한번 알려주면 끝인데 자네는 매번 이래 . 죄송하지만 팀장님한테 확인받고 한 작업입니다 . 뭐 그럼 내가 일부러 거야 . 그건 . 그런 식으로 일할거면 회사 관둬요 . 누가 보면 내가 괴롭힌다고 생각하겠어 그런 말씀은 지나치십니다 그럼 알아서 잘 하던가 . 눈치도 . 다음부터는 다시 확인하겠습니다 . 팀장님도 말씀 가려서 .
Example 3: 씨 이번에 또 틀렸잖아 . 죄송합니다 . 다시 확인하겠습니다 . 동기인 씨는 한번 알려주면 끝인데 자네는 매번 알려줘도 이래 . 죄송하지만 팀장님한테 확인받고 한 작업입니다 . 뭐 그럼 내가 일부러 그런다는 거야 . 그건 아니지만 . 그런 식으로 일할거면 회 관둬요 . 누가 보면 내가 괴롭힌다고 생각하겠어 그런 말씀은 지나치십니다 . 그럼 알아서 잘 하던가 . 눈치도 없이 . 다음부터는 다시 확인하겠습니다 . 팀장님도 말씀 가려서 해주세요 .
Example 4: 씨 이번에 또 틀렸잖아 . 죄송합니다 . 다시 확인하겠습니다 . 동기인 씨는 한번 알려주면 끝인데 매번 알려줘도 죄송하지만 팀장님한테 확인받고 한 작업입니다 . 뭐 그럼 내가 일부러 그런다는 거야 그건 아니지만 . 그런 식으로 일할거면 회사 관둬요 누가 보면 내가 괴롭힌다고 생각하겠어 그런 말씀은 지나치십니다 . 알아서 잘 하던가 . 눈치도 없이 . 다음부터는 다시 확인하겠습니다 . 팀장님도 가려서 해주세요 .
Example 5: 씨 씨 씨 씨 씨 씨 씨 씨 이번에 회 씨 또 씨 회 씨 씨 씨 틀렸잖아 회 씨 씨 씨 씨 씨 . 씨 씨 죄송합니다 회 씨 . 씨 씨 씨 씨 회 씨 회 다시 확인하겠습니다 씨 씨 씨 . 씨 동기인 씨 씨 씨는 회 한번 씨 알려주면 회 씨 씨 씨 씨 회 씨 끝인데 씨 자네는 회 씨 씨 씨 씨 씨 씨 씨 매번 씨 씨 씨 회 회 씨 씨 씨 씨 회 회 씨 씨 회 씨 씨 알려줘도 씨 회 씨 씨 씨 이래 씨 씨 씨 씨 씨 씨 . 씨 회 씨 죄송하지만 씨 씨 씨 회 회 씨 씨 회 씨 씨 씨 씨 씨 씨 씨 회 씨 씨 씨 회 씨 씨 회 씨 씨 씨 팀장님한테 씨 확인받고 회 씨 씨 씨 씨 회 씨 씨 씨 씨 회 씨 회 회 회 씨 회 씨 한 작업입니다 회 . 회 씨 씨 씨 뭐 회 씨 씨 그럼 씨 회 씨 씨 내가 씨 회 일부러 씨 씨 씨 회 씨 그런다는 거야 씨 씨 씨 씨 씨 회 씨 . 회 회 그건 회 아니지만 회 . 씨 씨 씨 씨 회 씨 회 씨 씨 씨 그런 씨 회 식으로 일할거면 씨 회 회사 씨 회 씨 씨 회 씨 씨 관둬요 씨 씨 씨 씨 씨 씨 씨 씨 씨 씨 씨 회 . 씨 씨 씨 누가 씨 씨 회 보면 씨 회 씨 내가 씨 씨 씨 회 씨 씨 씨 괴롭힌다고 씨 생각하겠어 회 씨 씨 씨 씨 그런 회 씨 씨 씨 회 말씀은 지나치십니다 씨 회 씨 씨 회 씨 씨 씨 씨 회 씨 . 그럼 씨 씨 회 씨 씨 씨 알아서 씨 잘 회 씨 회 하던가 . 눈치도 씨 씨 회 씨 회 없이 씨 씨 씨 . 회 씨 씨 회 씨 회 다음부터는 회 씨 회 씨 씨 씨 씨 씨 회 회 회 회 회 씨 회 다시 회 씨 씨 회 씨 확인하겠습니다 씨 씨 회 회 씨 씨 씨 씨 씨 회 회 회 씨 씨 회 씨 씨 씨 회 씨 씨 . 회 씨 회 씨 씨 팀장님도 씨 씨 씨 씨 씨 씨 회 씨 회 말씀 씨 씨 씨 씨 씨 씨 씨 가려서 씨 회 해주세요 씨 씨 .

AI HUB DATA

BACK TRANSLATION

Presentation are communication tools that can be used as demontrations, lectures, reports, and more. it is mostly presented before an audience.

Subject : BUSINESS MARKETING

Submit by : CLAUDIA ALVES

back translation

Back Translation : 합성 코퍼스를 만드는 기법 중 하나이다. Back Translation은 한 데이터의 한 문장을 가져와 다른 언어로 번역한 다음 다시 원래 언어로 번역하여 학습데이터의 양을 증가 시킨다. 즉 말뭉치 확장 기법의 일종이다

한 문장을 가져와 다른 언어로 번역한 다음 다시 원래 언어로 번역하는 기술

Thynk University

Start

ADDITIONAL DATA

Presentation are communication tools that can be used as demontrations, lectures, reports, and more. it is mostly presented before an audience.

Subject : BUSINESS MARKETING

Submit by : CLAUDIA ALVES