< DLTHON_NLP >

DKTC 다중분류 프로젝트: 한국어 위협 대화 데이터셋 활용

2024.01.11

팀 이름: 바른말 고운말

목차

DKTC 다중 분류 프로젝트

- 프로젝트 소개
- Data

진행 과정

- Exploratory Data Analysis
- Data Preprocessing
- Model Training
  - Machine Learning
  - Deep Learning
  - Transformer
- Model Evaluation

결론

- Conclusion
- References

DKTC 다중분류 프로젝트(DKTC Multi-Classification Project)

진행 일시

- **2024/01/10 ~ 2024/01/12**

목적

- **DKTC (Dataset of Korean Threatening Conversations)**를 활용하여 한국어 위협 대화에 대한 효과적인 다중 분류 모델을 구축하여 다양한 위협적인 대화에 대한 대응할 수 있는 솔루션을 구축 제공하고자 함

바른말 고운말 팀 전략

- 머신러닝과 딥러닝 기술을 적용하여 다중 분류 모델을 구축

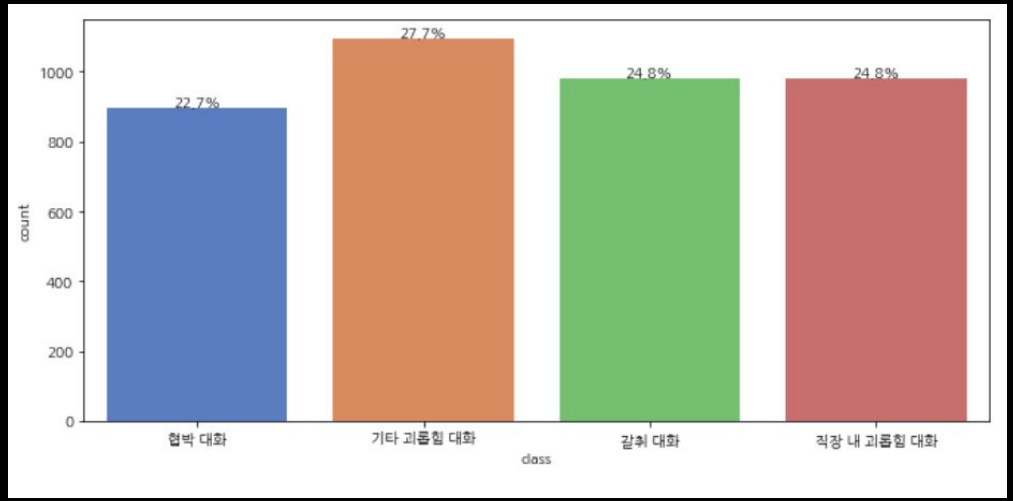
Data

- **TUNiB에서 제공된 DKTC (Dataset of Korean Threatening Conversations)**
- **멀티턴 대화 형식으로 구성되어 있으며, 3가지 파일(train, test, submission)로 이루어져 있음**
- **총 4가지 클래스로 세분화되어 있음**

class	conversation
협박 대화	지금 너 스스로를 죽여달라고 애원하는 것인가?\n 아닙니다. 죄송합니다.\n 죽을 ...
협박 대화	길동경찰서입니다.\n9시 40분 마트에 폭발물을 설치할거다.\n네?\n똑바로 들어 ...
기타 괴롭힘 대화	너 되게 귀여운거 알지? 나보다 작은 남자는 첨봤어.\n그만해. 니들 놀리는거 재미...
갈취 대화	어이 거기\n예??\n너 말이야 너. 이리 오라고\n무슨 일.\n너 옷 좋아보인다?...
갈취 대화	저기요 혹시 날이 너무 뜨겁잖아요? 저희 회사에서 이 선크림 파는데 한 번 손등에 ...
직장 내 괴롭힘 대화	나 이틀뒤에 가나다 음식점 예약좀 해줘. 저녁7시로.\n가나다 음식점이요.?\n응....

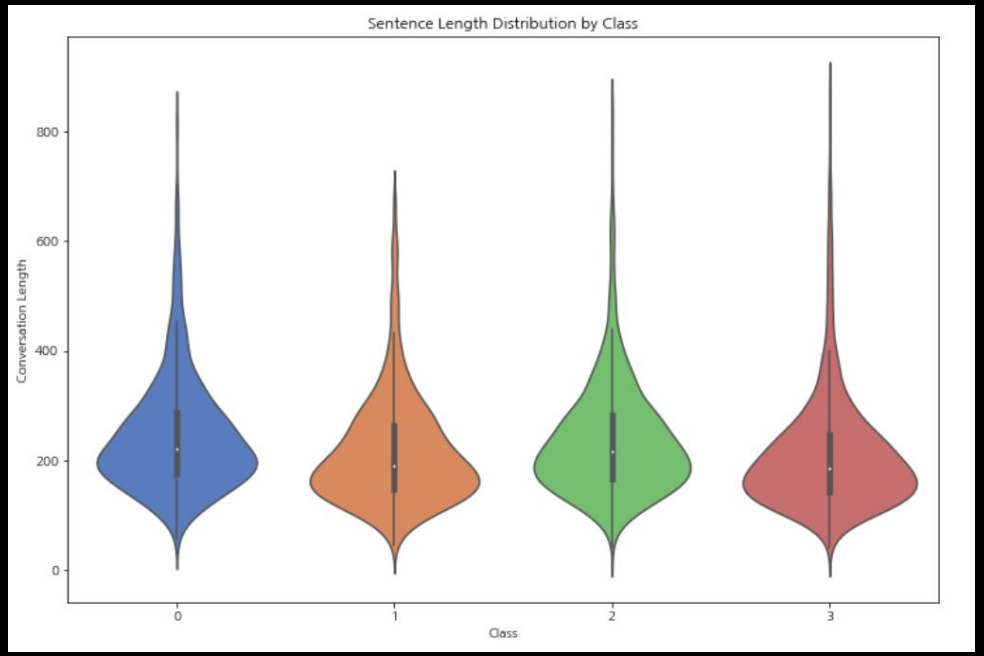
클래스	Class No.	# Training	# Test
협박	00	896	100
갈취	01	981	100
직장 내 괴롭힘	02	979	100
기타 괴롭힘	03	1,094	100

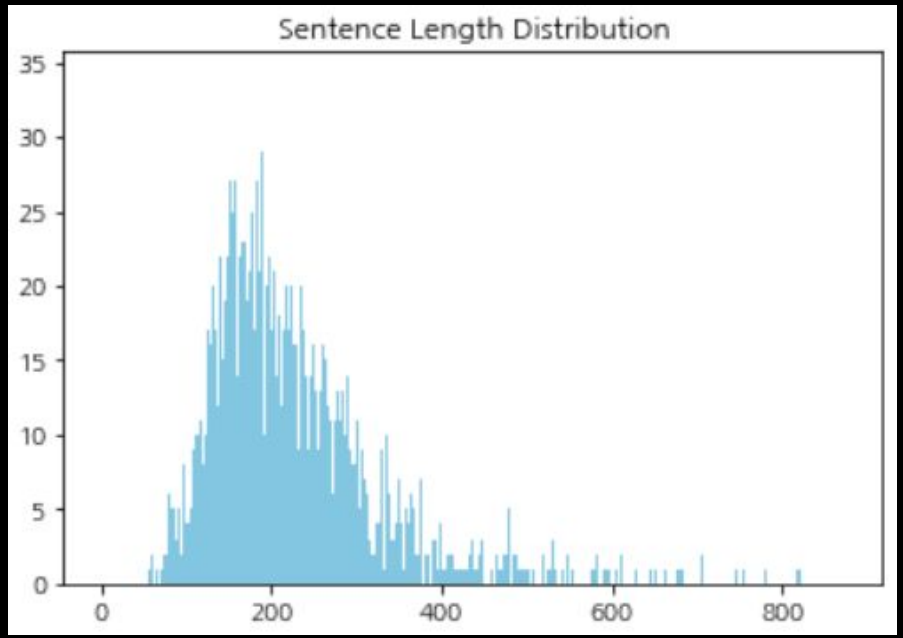
Exploratory Data Analysis



- 협박 대화 : **22.7 %**, 기타 괴롭힘 대화 : **27.7%**, 갈취 대화 : **24.8%**, 직장 내 괴롭힘 대화 : **24.8%**
- 클래스 간의 약간의 차이가 있기는 하나, 큰 차이를 보이지 않음
- 데이터불균형에 대해 크게 고려하지 않기로 함

Exploratory Data Analysis





- 클래스 별로 문장 길이 분포를 확인한 결과, 크게 차이가 나지 않는 걸로 보임
- 문장 최단 길이 : **41**, 문장 최장 길이: **874**, 문장 평균 길이: **226**

Data Preprocessing

- 문장 내 구두점 앞 뒤로 공백을 추가
- 연속된 여러 개의 공백을 하나의 공백으로 축소
- 한글, 영문자와 구두점(. ? ! ,) 이외의 모든 문자를 공백으로 대체

0	지금 너 스스로를 죽여달라고 애원하는 것인가?₩n 아닙니다. 죄송합니다.₩n 죽을 ...
1	길동경찰서입니다.₩n9시 40분 마트에 폭발물을 설치할거다.₩n네?₩n옥바로 들어 ...
2	너 되게 귀여운거 알지? 나보다 작은 남자는 첨봤어.₩n그만해. 니들 놀리는거 재미...
3	어이 거기₩n예??₩n너 말이야 너. 이리 오라고₩n무슨 일.₩n너 옷 좋아보인다?...
4	저기요 혹시 날이 너무 뜨겁잖아요? 저희 회사에서 이 선크림 파는데 한 번 손등에 ...

['지금 너 스스로를 죽여달라고 애원하는 것인가 ? 아닙니다 . 죄송합니다 . 죽을 거면 혼자 죽지 우리까지 사건에 휘말리게 해 ? 진짜 죽여버리고 싶게 . 정말 잘못했습니다 . 너가 선택해 . 너가 죽을래 네 가족을 죽여줄까 . 죄송합니다 . 정말 잘못했습니다 . 너에게는 선택권이 없어 . 선택 못한다면 너와 네 가족까지 모조리 죽여버릴거야 . 선택 못하겠습니다 . 한번만 도와주세요 . 그냥 다 죽여버려야겠군 . 이의 없지 ? 제발 도와주세요 .',

Tokenization

- **OKT, Twitter, SPM(SentencePiece)**을 사용하여 비교해 본 결과, **SPM**이 가장 우수한 성능을 보여, **SPM**을 사용하여 토큰화를 진행함

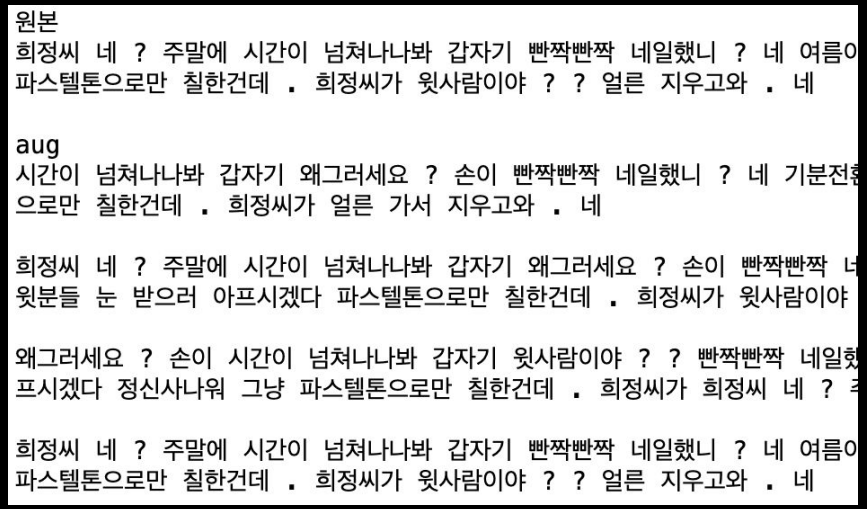
- **SGDClassifier** 기준 결과 비교

OKT	Twitter	SPM
0.285	0.285	**0.3075**

Data Augmentation

- 노이즈를 통한 일반화를 augmentation 사용하여 overfitting 방지

- 사용한 augmentation 방법
  - RS(Random swap): Randomly swap two words in the sentence n times
  - RD(Random deletion): Randomly delete words from the sentence with probability p
  - `num_aug = 5` 로 설정, 3846 -> 18958 (중복값 제거)로 데이터 수를 늘림



[출처 : KorEDA/README.md at master]

Model: Machine Learning

- 사용된 머신러닝 모델은 다음과 같음
  - **SGDClassifier**
  - **LinearSVC**
  - **VotingClassifier(Ensemble)**
  - **MultinomialNB**
  - **LogisticRegression**
  - **ComplementNB**
  - **DecisionTreeClassifier**
  - **RandomForestClassifier**
  - **GradientBoostingClassifier**
  - **LGBMClassifier**
  - **XGBClassifier**
    - \+ **augmentation 적용**

	SGD Classifier	LinearSVC	Ensemble
Validation_acc	0.858	0.858	0.862
Submission_acc	0.815	0.81	0.81
Aug적용 Submission_acc	0.82	0.815	0.81

* Ensemble: VotingClassifier(
estimators=[('CNB', ComplementNB()), ('SGD', SGDClassifier(random_state=42)), ('SVC', LinearSVC(random_state=42))])

Model: Deep Learning

- 사용한 딥러닝 모델은 다음과 같음

- SimpleRNN
- LSTM
- Bi-directional LSTM
- GRU

- + Augmentation

	validation_acc	submission_acc
Simple RNN	0.2790	0.3
LSTM	0.5373	0.54
bi-directional LSTM	0.5838	0.5675
GRU	0.5557	0.56

+Augmentation	validation_acc	submission_acc
aug + Simple RNN	0.4083	0.345
aug + LSTM	0.8294	0.7325
aug + bi-directional LSTM	0.8313	0.755
aug + GRU	0.7547	0.6275

Model: Transformer

- 사전 훈련된 언어 모델을 사용함

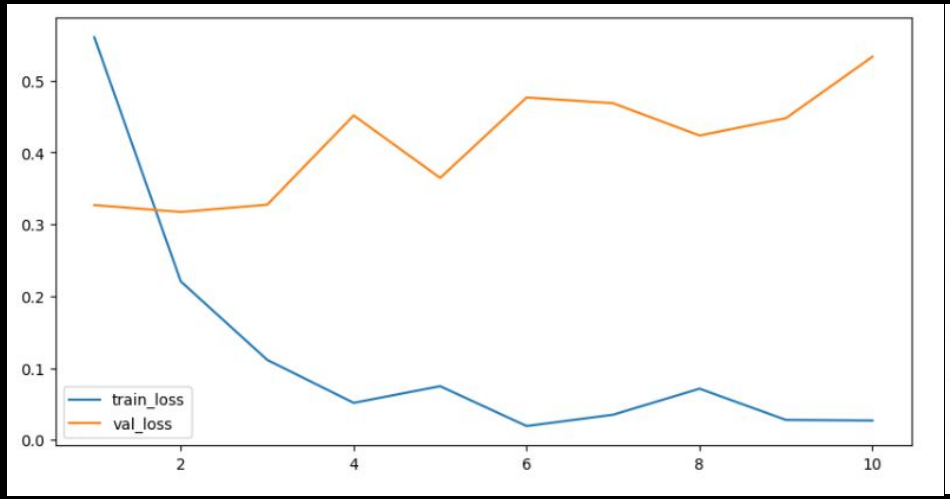
  - 한국어 Bert 계열 (accuracy 기준)
    - *klue/bert-base : 90%*
    - *kykim/bert-kor-base : 92.25%*
    - *monologg/kobigbird-bert-base:91.5%*
    - Bert 계열의 경우 토큰화로 BertTokenizerFast를 사용함

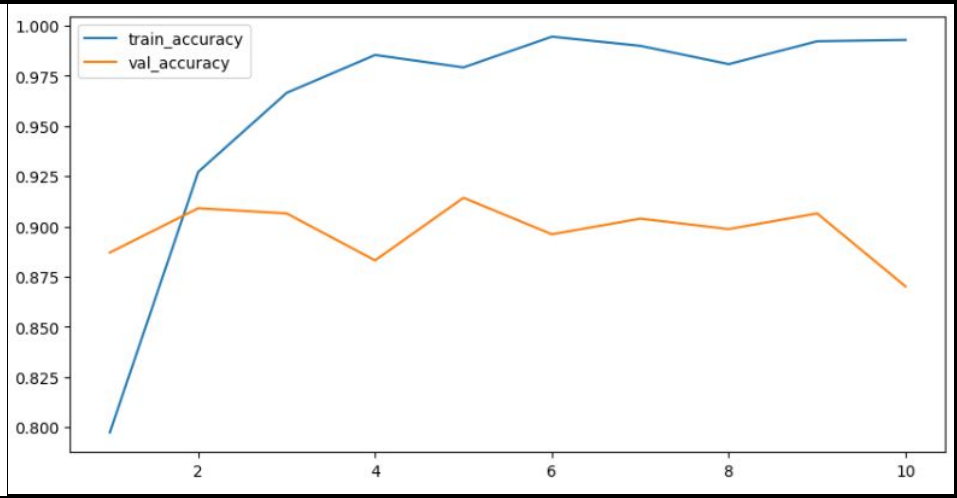
  - Bert 앙상블
    - 앙상블은 각 모델의 softmax결과를 평균

  - GPT 계열
    - *GPT-3.5 fine-tuning : 86.5%*
    - GPT의 경우 우수한 성능을 믿고 추가적인 전처리를 하지 않음
    - 비용과 시간의 문제로 400개의 문장만 넣음

Model: Transformer

- kykim/bert-kor-base : 92.25% 의 loss , accuracy 그래프 -> 과적합 의심





Model Evaluation

- 머신러닝 모델 중에서는 **aug + SGD Classifier(82%)**로 가장 높은 성능을 보임
- 딥러닝 모델 중에서는 **aug+bi_LSTM(75.5%)**로 가장 높은 성능을 보임
- **Transformer** 모델 중에서는 **kykim/bert-kor-base(92.25%)**로 가장 높은 성능을 보이며, 실험 모델 중에서 최고 성능을 보임

Team Name	Accuracy Score
바른말 고운말 simpleRNN	0.3
바른말 고운말 aug_simpleRNN	0.345
바른말 고운말 LSTM	0.54
바른말 고운말 GRU	0.56
바른말 고운말 bi_LSTM	0.5675
바른말 고운말 aug_GRU	0.6275
바른말 고운말 aug+LSTM	0.6825
바른말 고운말 aug_LSTM	0.7325
바른말 고운말 aug_bi_LSTM	0.755
바른말 고운말_ML2	0.81
바른말 고운말_MLensemble	0.81
바른말 고운말 ML	0.815
바른말 고운말 aug_ML2	0.815
바른말 고운말 aug_ensemble	0.8175
바른말 고운말 aug_ML	0.82
바른말 고운말 gpt3.5-ft	0.865
바른말 고운말 gpt3.5-ft	0.865
바른말 고운말 klue/bert-base	0.9
바른말 고운말 bert augment	0.9125
바른말 고운말	0.915
바른말 고운말 kobigbird	0.915
바른말 고운말 bert ensemble	0.92
바른말 고운말 kykim/bert-kor-base	0.9225

Conclusion

- 한국어 위협 대화에 대한 효과적인 다중 분류 모델을 구축하고자 다양한 모델을 통해 실험한 결과, **kykim/bert-kor-base(92.25%)** 모델이 가장 높은 성능을 보였음
- 이는 다양한 위협적인 대화에 대한 대응할 수 있는 솔루션으로 채택 될 수 있음
- **Augmentation Data**를 활용한 경우, 모델에 따라 성능 향상을 확인 할 수 있어 데이터 확보의 중요성을 강조 할 수 있음
- 본 프로젝트는 **Evaluation metrics** 선정에 있어 **accuracy**만 확인한 아쉬운 점이 있음
- 향후에는 다양한 평가 지표를 활용하여 모델의 성능을 보다 정확하게 평가하고, 데이터 확보와 모델 튜닝을 통해 다양한 상황에서 적용 가능하며 강건한 모델을 개발하는 것을 기대해 볼 수 있음

Reference

- https://github.com/catSirup/KorEDA/blob/master
- https://velog.io/@jaehyeong/Fine-tuning-Bert-using-Transformers-and-TensorFlow
- https://huggingface.co/models?pipeline_tag=fill-mask&language=ko&sort=likes&search=bert

들어주셔서 감사합니다!