Robust Sentimental prediction using Bert-CNN based word- and character- level embedding

Sentimeter, 박형준

1. Introduction

- Bert와 같은 가장 발전된 자연어 분석 모델들은 단어 단위의 주요 latent feature를 추출에 집중하고 있다.

- 하지만, 트위터 글 및 영화 리뷰 문장은 오타, 기호 및 신용어로 구성된 경우가 많기에 이러한 단어들의 의미를 정확히 분석하는 것에 한계가 있다. (OOV 문제)

- 이러한 한계에 대응하기 위해서 문자 단위의 embedding을 추가 정보로 주면 의미를 파악하는데 도움이 될 수 있다고 판단하여 본 프로젝트에서는 CNN기반의 문자 단위의 embedding을 BERT의 단어 단위 embedding과 함께 사용한다.

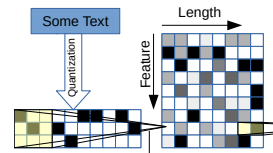
1. Novelty

* Embedding 단계에서 Word-level 및 character-level embedding 방식을 동시에 적용하여 영화 리뷰에서 주요한 latent feature를 추출한다.
* Fine-tuning 단계에서 drop out이 적용된 여러 classifier을 앙상블로 구성하여 강건한 감정 예측을 수행한다.

1. Methodology
   1. Preprocessing

- Bert model: 알파벳 소문자로 변환 전처리 후 Bert tokenizer를 이용하여 문장내 단어 단위 토큰화 수행

- CNN model: 알파벳 소문자로 변환 전처리 후 각 문자를 one-hot encoding을 수행한다. 문장 내의 문자들을 순차적으로 쌓아서 2차원의 matrix(character\_len X sentence\_len)로 변환된다.



* 1. Model Fitting

- Bert model: hugging face에서 제공하는 Bert pretrained model인 ‘bert-base-uncased’ 사용

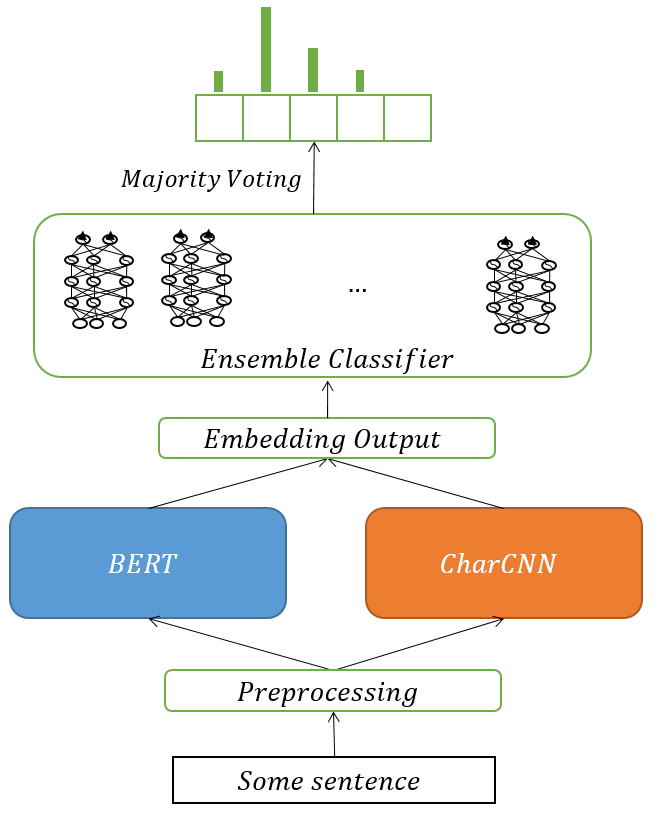
- CNN model: 프랑스어 제품 리뷰로 학습된 Character-level CNN pretrained model 사용

- Bert model에서 embedding한 단어 단위 latent feature (768 dim)과 CNN model에서 embedding한 문자 단위 latent feature (1024 dim)을 결합하여 최종 단어-문자 latent feature vector (1792 dim)을 embedding의 결과로 도출함

- embedding 결과를 ensemble classifier set에 속한 각 classifier에게 전달하여 각각의 classifier의 loss를 평균하고 이를 최소화 시키도록 파라미터를 조정하는 fine-tuning 과정 수행

- 예측시에는 ensemble classifier 각각에서 최대로 예측한 label set에서 majority voting을 통해 감정 예측을 수행함

* 1. Model structure



1. Performance

Score: 0.64733