**모델별 성능 표**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CounterVectorizer 수정 | N-gram | TF-IDF | 형태소 분석 입력 | 불용어 처리 | 성능 | 모델이름 |
| O | (1,1) | X | X | X | 80.16 | 기준모델 |
| O | (1,2) | X | X | X | 80.8 | 모딜2.1 |
| O | (1,3) | X | X | X | 80.16 | 모델2.2 |
| O | (1,5) | X | X | X | 79.53 | 모델2.3 |
| O | X | O | X | X | 80.88 | 모델3 |
| O | (1,2) | O | X | X | 82.26 | 모델4 |
| O | X | X | O | X | 82.3 | 모델5 |
| X | X | X | O | O | 79.75 | 모델6 |
| O | (1,1) | X | O | X | 82.30 | 모델7 |
| O | (1,2) | X | O | X | 84.03 | 모델8 |
| O | (1,3) | X | O | X | 85.11 | 모델9 |
| O | (1,1) | O | X | X | 80.88 | 모델10 |
| O | (1,2) | O | X | X | 82.26 | 모델11(모델4와동일) |
| O | (1,3) | O | X | X | 81.85 | 모델12 |
| O | (1,1) | O | O | X | 83.76 | 모델13 |
| O | (1,2) | O | O | X | 85.98 | 모델14 |
| O | (1,3) | O | O | X | 86.20 | 모델15 |
| X | (1,2) | O | O | O | 82.45 | 모델16 |

**각 실험에 대한 분석**

1. 모델 8

본 모델은 허깅페이스에서 제공하는 한국어 감정 분류 데이터셋 중 하나인 sepidmnorozy/Korean\_sentiment 데이터셋을 사용한다. 모델 학습, 검정, 평가를 위해 train, validation, test 로 세트를 구성하였다. 텍스트 데이터는 Bag-of-words를 이용하여 벡터화하였고, 이를 위해 사이킷런에서 제공하는 CountVectorizer 클래스를 사용하였다. 학습을 시키는 모델은 서포트 벡터 머신을 사용하였으며, 5-fold 교차 검증을 통해 모델의 성능을 평가하였다. 실험의 시간적 효율성을 위해 한 번 학습된 모델을 사용하여 test 셋을 평가하였다.(그리드 서치 사용 X)

다만 실험 중 CountVectorizer를 사용했을 때 일부 단어들이 사라지는 현상이 있어, 악센트 제거 및 캐릭터 노말라이제이션을 수행하는 strip\_accent 파라미터의 값을 unicode로, 정규표현식으론 토큰 패턴을 설정해줌으로 한국어 단어 소실 문제를 해결하였다.

또한 본 모델은 단어 예측에 있어서 문장을 1개, 2개의 단어로 나누는 unigram, bigram feature를 사용하였으며 konlpy에서 제공하는 mecab를 활용하여 형태소 단위로 토크나이징된 텍스트를 입력 데이터로 사용하였다. 정확도는 소수점 두 자리수까지 출력한 결과 84.03의 정답률을 확인할 수 있었다.

1. 모델 9

해당 모델은 모델 8과 동일한 환경과 파라미터 아래에서 실험을 수행했지만, 단어 예측에 있어 모델 8의 unigram, bigram 이외에도 trigram의 feature를 추가적으로 사용하였다. 모델의 정확도를 소수 둘째자리까지 출력한 결과 85.11의 값을 가졌다.

1. 모델 10

모델 10의 경우 앞선 모델들과 달리 mecab을 활용한 형태소 분석을 거치지 않고 주어진 텍스트를 모델의 입력값으로 사용해 분석을 수행하였다. 또한 단어 예측에 있어서 feature로는 unigram feature만을 사용했다는 차이점이 있다. 또한 기존의 bag-of-words 방법을 TF-IDF 방식으로 바꾸었다는 차이점도 있다. 이외의 모델, 노말라이제이션, 평가 방법 등은 앞선 실험들과 동일하게 가져갔으며, 정답률은 소수 둘째자리까지 출력한 결과 80.88로, 앞선 모델과 비교하였을 때 토크나이징이 분석의 정답률 향상에 있어 유의미한 영향을 끼침을 확인하게 한다.

1. 모델 12

모델 12는 앞서 모델 10과 동일한 환경 및 파라미터를 사용하지만, feature에 있어 bigram, trigram을 추가로 사용했다는 차이점이 있다. 이에 정답률은 81.85로, 모델 10과 비교하였을 때 미미한 상승을 보였다.

1. 모델 13

모델 13은 앞서 모델 12와 동일하게 TF-IDF feature, 노말라이제이션 및 악센트 제거, 모델 등의 환경과 파라미터를 사용했지만 N-gram feature에 있어 unigram feature만을 사용했다는 점, 그리고 모델 8,9에서 사용한, mecab을 활용한 형태소 분석을 수행했다는 점에 있어 차이를 보인다. 정답률은 83.76으로 형태소 토크나이징 외의 환경이 동일한 모델 10과 비교했을때 정답률 향상으로 보아 토크나이징이 정답률 향상에 유의미한 영향을 끼침을 또다시 확인할 수 있다.

1. 모델 14

모델 14의 경우 앞서 모델 13에 bigram feature를 추가로 사용하였다. 이에 정답률이 85.98로 소폭 상승하는 것으로 보아 N-gram feature 선택에 있어 unigram 단독 사용보다 bigram도 함께 사용하는 것이 정답률 향상에 있어 유의미한 결과를 낳는다는 것을 확인할 수 있다.

1. 모델 15

모델 15는 앞서 모델 14에 trigram feature를 추가로 사용한 실험이다. 이에 정답률이 86.20으로, 지금까지 실험한 모든 경우 중 정답률이 가장 높은 것으로 미루어 보았을 때, N-gram feature의 경우 토크나이징을 수행한다면 bag-of-words나 TF-IDF 자질과 상관없이 많이 사용할수록 정답률이 높아짐을 확인할 수 있다.(모델 7,8,9 / 모델 13,14,15 비교)

1. 모델 16

모델 16은 형태소 토크나이징을 거치고, 형태소 중 조사를 불용어 처리했다. N-gram feature 중 unigram, bigram feature와 TF-IDF feature를 사용하였다. 정답률은 82.45로, CounterVectorizer의 토큰 소실을 개선한 모델인 모델 14와 비교했을 때 성능이 3.53p 하락했다.

**모델 선택**

최종 모델 하나를 선택해야 한다면, 모델 15를 선택할 것이다. 머신러닝 예측에 있어서 가장 중요한 것은 성능이다. 물론 F1 score, Recall, Precision 등 모델의 성능을 평가하는 산식을 여러 가지 사용할 경우 모두 고려해 성능이 가장 좋은 모델을 택해야 하겠지만, 본 실험의 경우 Accuracy만을 성능 평가 산식으로 사용했기 때문에, Accuracy가 가장 높은 모델 15를 선택하는 것이 타당하다고 생각한다.

**추가 실험**

모델 15의 환경을 동일하게 가져가되, N-gram feature를 (1,4), (1,5)로 수정하여 정답률을 확인해보고 싶다. 모델별 성능 표를 보면 모델 13,14,15가 N-gram feature의 개수만을 늘리면서 정답률을 꾸준히 향상시켰고, 모델 15의 경우 가장 높은 성능을 보임을 알 수 있다. 다만 본 성능 표가 제시하는 파라미터 수정 범위 내에서 모델 15가 최선의 값을 만든다는 것을 보장하기는 어렵다. 왜냐하면 기준모델, 모델2.1, 모델2.2 를 비교함으로 우리는 countervectorizer를 수정하고, TF-IDF, 불용어처리, 형태소 토크나이징을 사용하지 않을 경우 N-gram feature는 (1,2)의 값이 최적임을 알 수 있었기 때문이며, 이보다 더 늘릴 시 정답률이 하락하는 것을 보았기 때문이다. 이와 같은 방식으로, 모델 15의 N-gram feature (1,3)의 값이 최적인지를 확인하는 과정이 필요하다고 생각한다. 만일 (1,4), (1,5)로 N-gram feature 값을 수정하였을 때 정답률이 (1,3)의 경우보다 낮을 시, N-gram feature는 (1,3)으로 설정하는 것이 가장 좋을 것이다. 실제로 실험해본 결과, (1,4)의 경우 정답률이 모델 15의 것과 동일한 86.20, (1,5)의 경우 85.98로 하락하는 것을 확인했다. 이 과정을 통해 모델 15의 파라미터 값이 현 상황에선 최선의 것임을 보장할 수 있다.