NLP Assignment 2

Naive Bayes sentiment classification의 inference가 제대로 수행되려면 대부분의 머신러닝, 딥러닝 분야에서는 Data Processing과정과 Inference의 analysis가 중요하다고 개인적으로 생각합니다.

따라서, 이 보고서에는 위 2가지에 대해 설명하겠습니다.

코드는 py, ipynb 1개의 코드를 2가지 형식으로 올렸습니다.(혹시 몰라서 2개로 올렸습니다!)

Data Processing

1. 구두점 제거
    a. Noise로 간주되는 구두점을 제거하였습니다.
    b. 이는, text의 실제 의미와 관계가 없는 요소로 판단하였으며, 토큰화의 일관성을 향상시켰습니다.

2. 소문자 변환
    a. 대소문자의 구분으로 똑같은 단어이지만, 대문자와 소문자 때문에 다르게 해석될 위험이 있어 소문자로 통일된 형식으로 변환하여 일관성을 부여했습니다.
    b. 이는, Dataset의 어휘 크기를 줄여줄 수 있습니다.

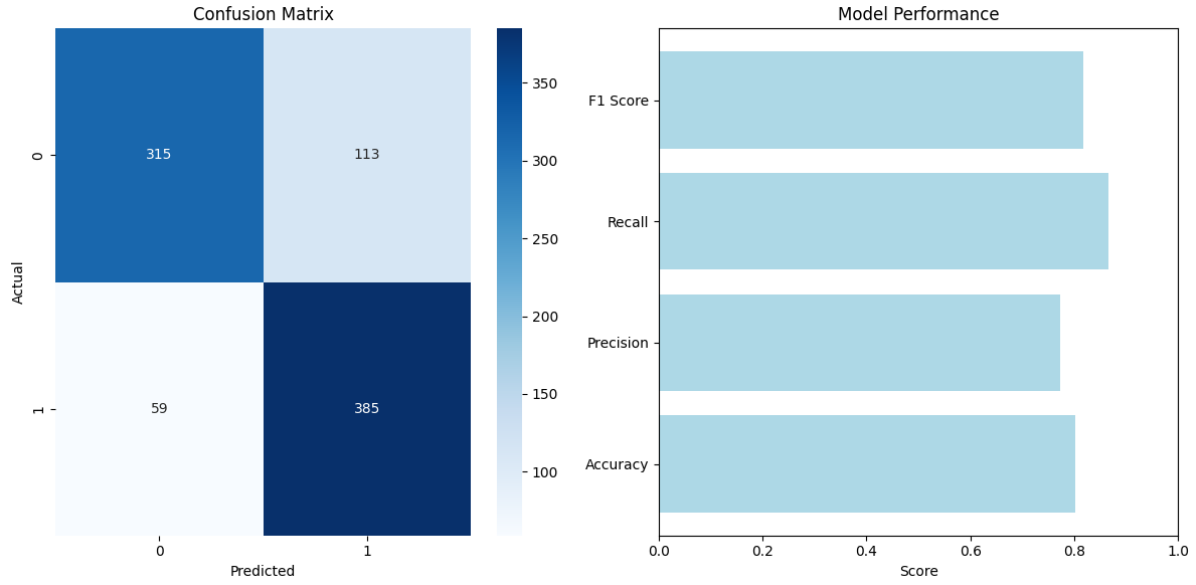
3. 토큰화
    a. text를 작은 단위로 분할하여 구조화할 수 있게끔 했습니다.
    b. Likelihood를 계산하기 위해 토큰화를 진행을 하기도 했습니다.

4. 불용어 제거
    a. 문법적으로는 중요한 단어이지만, text analysis에서는 큰 영향을 미치지 않는다고 생각했습니다.

b. 이는, 핵심 단어에 집중할 수 있도록 했습니다.

5. 표제어 추출
   a. 단어의 다양한 형태를 하나의 표제어로 통일하게 되면, 어휘의 일관성이 향상 된다고 합니다. 이를 통해, text를 더 효율적으로 분석할 수 있게 되었습니다.
   b. 어휘의 다양성을 줄였습니다. 이는 data의 dimension을 낮추기 때문에 inference과정에서 더 효율적입니다.

결과 분석



이 모델은 80.28%의 Accuracy를 가지고 있으며, F1 Score를 통해 Precision과 Recall사이의 균형이 좋은 편임을 알 수 있습니다.(F1 Score : 0.8174)

그러나, Positive 예측은 잘하고 있지만, 여전히 약간의 오차(113개의 False Negative 및 False Positive)를 가지고 있음을 볼 수 있습니다. 이는 Recall이 높고 Precision이 낮다는 것을 보여줍니다. (Recall : 0.8671, Precision : 0.7731)

따라서, Positve class를 정확하게 파악하고 검출하는 능력이 우수합니다. 이는, 질병 진단, 스팸 메일 filtering 등에서 쓰일 수 있습니다.