로이터 뉴스 데이터셋을 이용한 전통적 머신 러닝 모델과 딥 러닝 모델의 성능 비교 분석

Seoyeon Kim

Abstract

뉴스 분류는 대량의 디지털 정보를 효율적으로 처리하는 핵심 태스크로, 머신 러닝과 딥러닝 모델을 이용하여 자동화된 방식으로 주제 및 카테고리를 식별하고 분류한다. 이 연구는 이러한 모델들이 뉴스 분류에서 어떤 성능을 보이는지 분석한다.

1. Introduction

텍스트 분류(Text Classification)란 주어진 텍스트를 사전 정의된 클래스(Pre-defined Class)들로 분류하는 자연어 처리 태스크이다. 이 태스크는 자연어 처리 태스크 중 가장 기본이 되면서, 비즈니스 영역에서도 가장 수요가 높다. 특히 뉴스 분류는 디지털 정보가 급증하는 시대에 정보의 정확한 관리와 효율적인 검색, 그리고 사용자 맞춤형 콘텐츠 제공을 위해 필요한 태스크이다. 이 과정은 대규모 텍스트 데이터에서 특정 주제나 카테고리를 식별하고 분류하는 것을 포함한다. 이러한 태스크는 수동으로 처리하기에는 너무 방대하고 복잡하여, 자동화된 방법이 필수적이다.

머신 러닝과 딥러닝 모델은 이 분야에서 강력한 도구로 자리 잡았다. 이들은 대용량의 데이터로부터 패턴을 학습하고, 뉴스 기사와 같은 텍스트 데이터를 효과적으로 분류할 수 있는 능력을 가진다. 머신 러닝 모델은 전통적으로 통계적 방법과 알고리즘을 활용하여 데이터의 특성을 학습한다. 반면, 딥러닝 모델은 인공 신경망을 사용하여 보다 복잡하고 추상적인 데이터의 표현을 학습할 수 있다. 본 연구는 이러한 모델들이 뉴스 분류라는 구체적인 문제에 적용되었을 때 어느 정도의 성과를 거두는지를 다룬다.

2. Method

2.1. Dataset

본 연구에서 사용한 데이터는 로이터 뉴스 데이터로, 총 46개의 클래스로 구성되며 해당 뉴스가 어느 카테고리에 속하는지를 예측하기 위한 데이터이다. TensorFlow 데이터셋을 통해 제공받았으며 학습용으로 8,982개, 테스트용으로 2,246개의 뉴스를 사용하였다. 데이터셋 구축 과정에서 총 단어 개수에 제한을 두지 않은 경우와 5000개, 1000개로 제한한 경우 3가지로 나누어 각각의 경우에서 학습 모델의 성능이 어떻게 달라지는지 비교관찰하였다.

2.2. Models

인공신경망을 활용하지 않는 전통적인 머신러닝 모델 6개를 선정하여 뉴스 분류 태스크를 진행하였다. 6개의 모델은 각각 나이브 베이즈 분류기, 컴플리먼트 나이브 베이즈 분류기, 로지스틱 회귀 모델(최대 반복 횟수 3000), 선형 서포트 벡터 머신, 결정 트리(최대 깊이 10), 랜덤 포레스트(분류기 개수 5개)이다. 별도로 언급하지 않은 하이퍼파라미터는 기본값을 사용하였다.

딥러닝 모델의 경우 LSTM을 사용한 모델과 1차원 합성곱을 사용한 모델 2가지를 사용하였다. LSTM 모델은 임베딩 층(임베딩 차원=50, 최대 길이 300), LSTM 층(유닛=50), 드랍아웃 층(드랍아웃 비율=0.4), 배치 정규화 층, 밀집 연결층(유닛=60, 활성화 함수=relu), 그리고 최종 분류를 수행하는 밀집 연결층 순으로 구성된 모델이다. 1차원 합성곱 모델은 임베딩 층(임베딩 차원=50, 최대 길이 300), 1차원 합성곱 층(필터=30, 커널 크기=5, 활성화 함수=relu), 최대 풀링 층(풀링 크기=5), 전역 평균 풀링 층, 그리고 최종 분류를 수행하는 밀집 연결층 순으로 구성된 모델이다.

2.3. Metric

모델을 평가하는 지표로 정확도와 F1-스코어를 이용하였다.

3. Results

3.1. ML Models

전통적인 머신 러닝 모델에 대해 학습 후 평가를 진행한 결과 선형 서포트 벡터 머신이 모든 경우에 대해 가장 좋은 성능을 보였다(Figure 1, 2 참조). 단어 개수에 제한을 두지 않은 경우 82.95%의 성능을 보였다. 선형 서포트 벡터 머신에 대해 단어 개수 별로 F1-스코어를 비교한 결과 Figure 3에서 볼 수 있듯이 단어 개수에 제한을 두지 않은 경우 가장 좋은 성능을 보임을 알 수 있다.

3.2. DL Models

LSTM 모델에서 약 63%, 합성곱 모델에서 약 73%의 정확도를 보였다.

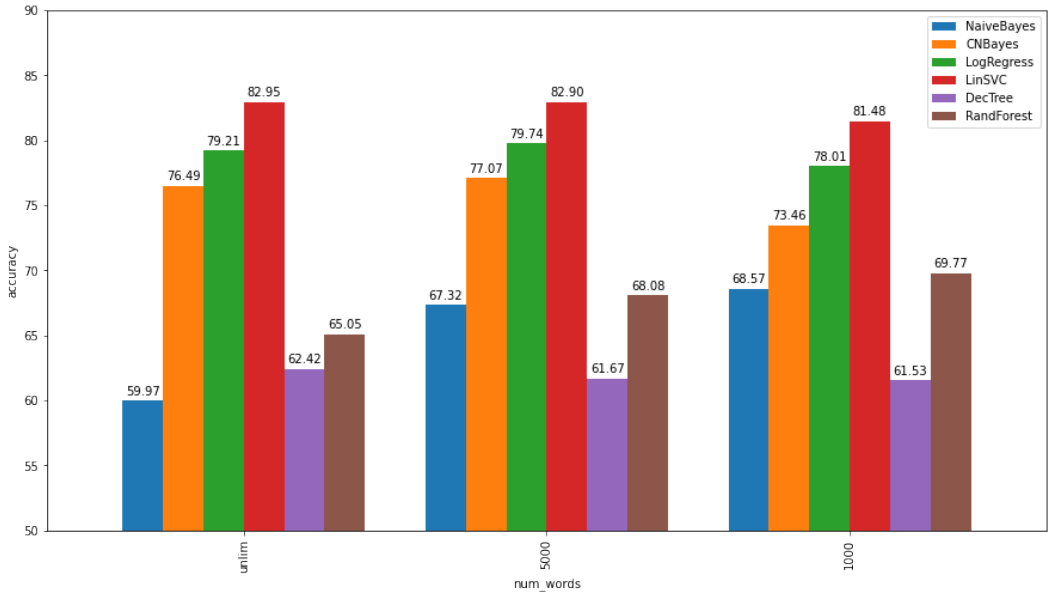

Figure 1. 단어 개수에 따른 ML 모델 별 정확도 비교

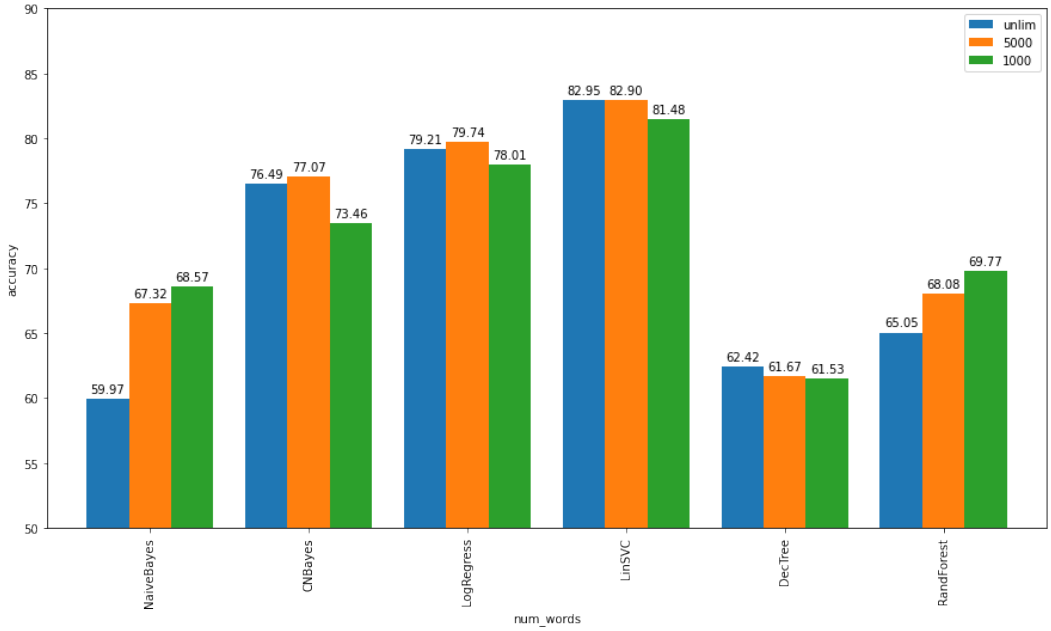

Figure 2. ML 모델에 따른 단어 개수 별 정확도 비교

4. Conclusion

딥러닝 모델이 선형 서포트 벡터 머신 모델보다 정확도가 떨어지는 것을 볼 수 있었다. 본 연구에서 사용된 데이터 개수는 약 10000개로 딥러닝 모델이 유용하게 사용되는 경우에 비해 훨씬 적은 데이터를 갖고 진행되었다. 데이터가 적은 경우 일반적인 머신러닝 모델이 학습 속도도 빠르고 성능도 뒤쳐지지 않거나 오히려 더 좋을 수 있다.

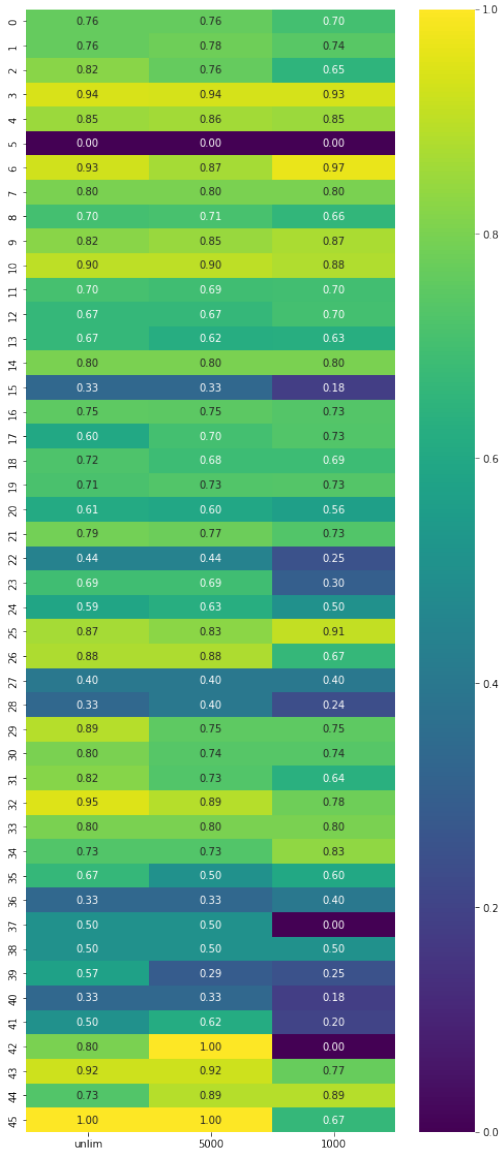

Figure 3. 선형 서포트 벡터 머신 모델의 단어 개수 별 F1-스코어 비교