JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 5, pp. 39-46, May 2021
https://doi.org/10.9708/jksci.2021.26.05.039

A Text Content Classification Using LSTM For Objective Category Classification

Young-Dan Noh*, Kyu-Cheol Cho*

*Student, Dept. of Computer Science, Inha Technical College, Incheon, Korea
*Professor, Dept. of Computer Science, Inha Technical College, Incheon, Korea

[Abstract]

AI is deeply applied to various algorithms that assists us, not only daily technologies like translator and Face ID, but also contributing to innumerable fields in industry, due to its dominance. In this research, we provide convenience through AI categorization, extracting the only data that users need, with objective classification, rather than verifying all data to find from the internet, where exists an immense number of contents. In this research, we propose a model using LSTM(Long-Short Term Memory Network), which stands out from text classification, and compare its performance with models of RNN(Recurrent Neural Network) and BiLSTM(Bidirectional LSTM), which is suitable structure for natural language processing. The performance of the three models is compared using measurements of accuracy, precision, and recall. As a result, the LSTM model appears to have the best performance. Therefore, in this research, text classification using LSTM is recommended.

▸ **Key words**: AI, contents, categorization, LSTM, natural language

[요 약]

인공지능은 현재 인공지능 번역기, 페이스 아이디와 같이 우리의 삶 다양한 곳에 적용되고 있으며 여러 가지 장점으로 많은 산업분야에서도 적용되고 있다. 본 연구는 매년 방대한 양의 콘텐츠들이 넘쳐나는 상황에서 인공지능을 적용한 카테고리 분류로 원하는 데이터를 추출함으로써 편의성을 제공한다. 본 연구에서는 텍스트 분류에서 두각을 나타내고 있는 LSTM(Long-Short Term Memory network)을 사용한 모델을 제안하며 자연어 처리에 적합한 구조를 가진 RNN(Recurrent Neural Network)과 BiLSTM(Bidirectional LSTM)을 사용한 모델과의 성능을 비교한다. 세 가지 모델의 성능비교는 뉴스 텍스트 데이터에 적용해 accuracy, precision, recall의 측정값을 사용해 비교하였고 그 결과 LSTM모델의 성능이 가장 우수한 것으로 나타났다. 따라서 본 연구에서는 LSTM을 사용한 텍스트 분류를 권장한다.

▸ **주제어**: 인공지능, 콘텐츠, 카테고리, LSTM, 자연어

• First Author: Young-Dan Noh, Corresponding Author: Kyu-Cheol Cho
*Young-Dan Noh (shdbwjd705270@gmail.com), Dept. of Computer Science, Inha Technical College
*Kyu-Cheol Cho (kccho@inhatc.ac.kr), Dept. of Computer Science, Inha Technical College
• Received: 2021. 01. 18, Revised: 2021. 04. 17, Accepted: 2021. 04. 20.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

AI기술은 현재 인공지능 번역기, 페이스 아이디와 같이 우리의 삶 다양한 곳에 이미 많이 적용되고 있다[4][5]. 인공지능의 장점은 빠른 처리능력, 비용절감, 자동화 등 다양하며 많은 산업분야에서도 적용되고 있는 상황이다.

AI기술을 활용하여 의료분야에서도 큰 성과를 보이고 있는데 그 예시로 AI 영상분석으로 자궁경부암 진단을 하는 알고리즘의 개발을 들 수 있다[6]. 이와 같이 의료분야에서 AI를 사용하면 진단의 정확도를 높일 수 있고 진단시간과 비용을 줄일 수 있다는 장점이 있다.

매년 방대한 양의 콘텐츠들이 넘쳐나는 상황에서 모든 데이터를 개인이 분류해 필요한 정보를 찾아내기보다는 인공지능을 적용한 카테고리 분류로 원하는 데이터를 추출함으로써 편의성을 제공한다.

본 연구에서는 다중 카테고리인 뉴스 데이터를 사용해 훈련을 진행한다. 현재 네이버는 연예인 인격권 보호와 사생활 피해 최소화를 위해 연예 뉴스 댓글 서비스를 폐지했다. 하지만 언론사들이 연예뉴스를 생활문화 섹션에 올리면서 또다시 악성댓글들이 달리고 있는 상황이다. 연예 기사를 생활문화 섹션으로 분류하게 되면 네이버 랭킹뉴스 순위에 손쉽게 올라갈 수 있다는 점을 이용한 것이라는 지적도 있다[7]. 만일 AI도입을 통한 객관적인 뉴스 카테고리 분류기가 있다면 연예뉴스 카테고리로 분류되어 악성댓글을 막을 수 있을 것이다.

카테고리 분류가 정확하지 않으면 뉴스를 보는 사용자 또한 혼란에 빠지기 쉽다. 사용자가 보려는 뉴스 카테고리에서 다른 카테고리에 해당하는 뉴스를 보게 된다면 원하는 뉴스만 다시 골라야 하는 사용자의 번거로움이 증가할 것이다. 따라서 본 연구는 객관적인 카테고리 분류기를 만드는 것을 목적으로 한다.

현재는 뉴스 데이터를 기반으로 텍스트 분류를 진행하였지만 카테고리 분류가 필요한 또 다른 상황에서도 활용될 수 있을 것으로 기대된다.

본 연구에서는 인공지능의 구현방법 중 하나인 딥러닝을 통해 한국어로 된 뉴스 텍스트를 분석하고 뉴스의 카테고리를 분류한다. 또한 단순히 형식에 맞는 CSV파일을 웹페이지에 업로드하면 자동으로 딥러닝을 통해 텍스트를 분석하고 분류해 예측 성공에 대한 차트를 제공한다.

본 연구의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 연구에 대한 기술한다. 3장에서는 LSTM(Long-Short Term Memory network)모델을 사용한 분석에 대해 기술한다. 4장에서는 RNN(Recurrent Neural Network), LSTM, BiLSTM(Bidirectional LSTM)모델을 사용한 분석에 대해 비교하며 사용자에게 제공되는 차트에 대해 기술하고 5장에서는 결론에 대해 기술한다.

II. Preliminaries

1. Related works

1.1 RNN(Recurrent Neural Network)

RNN은 과거의 학습결과를 현재 학습에 사용하는 인공신경망이다. 따라서 RNN의 경우 순서라는 측면을 고려할 수 있는 특징이 있어 시계열 데이터를 다루는 데 도움이 될 수 있다[8][9]. 시계열 데이터의 경우 '문장'과 같은 데이터를 예로 들 수 있다.

하지만 RNN은 한 정보와 그 정보를 사용하는 지점 사이의 격차가 커진다면 학습능력이 크게 저하된다는 장기의존성문제가 있다. 이런 RNN의 단점을 보완한 것이 바로 LSTM이다[9].

1.2 LSTM(Long-Short Term Memory network)

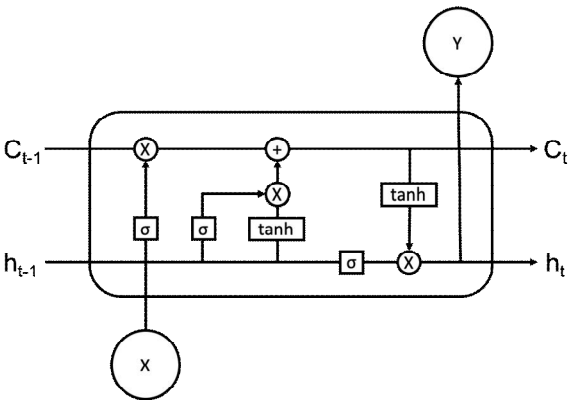

Fig. 1. LSTM cell

LSTM은 RNN의 장기의존성 문제를 해결한 인공 신경망이다. LSTM은 어떠한 정보는 기억하고 어떠한 정보는 잊는다. Fig. 1.은 LSTM의 구조를 나타낸다. 첫 번째로 sigmoid함수를 사용해 몇 퍼센트의 기억을 남길지 결정하고 두 번째로 sigmoid함수와 tanh함수를 사용해 기억할 새로운 정보를 결정한다. 세 번째로 잊어버리기로 결정한 것과 기억하기로 결정했던 것을 실제로 실천하고 마지막으로 무엇을 출력할지 결정한다[9][10].

LSTM은 내부의 정보가 지속되도록 순환구조를 가진 시계열 개념이 추가된 순환신경망으로 순환신경망은 텍스트 분류에서 두각을 나타내고 있다[11].

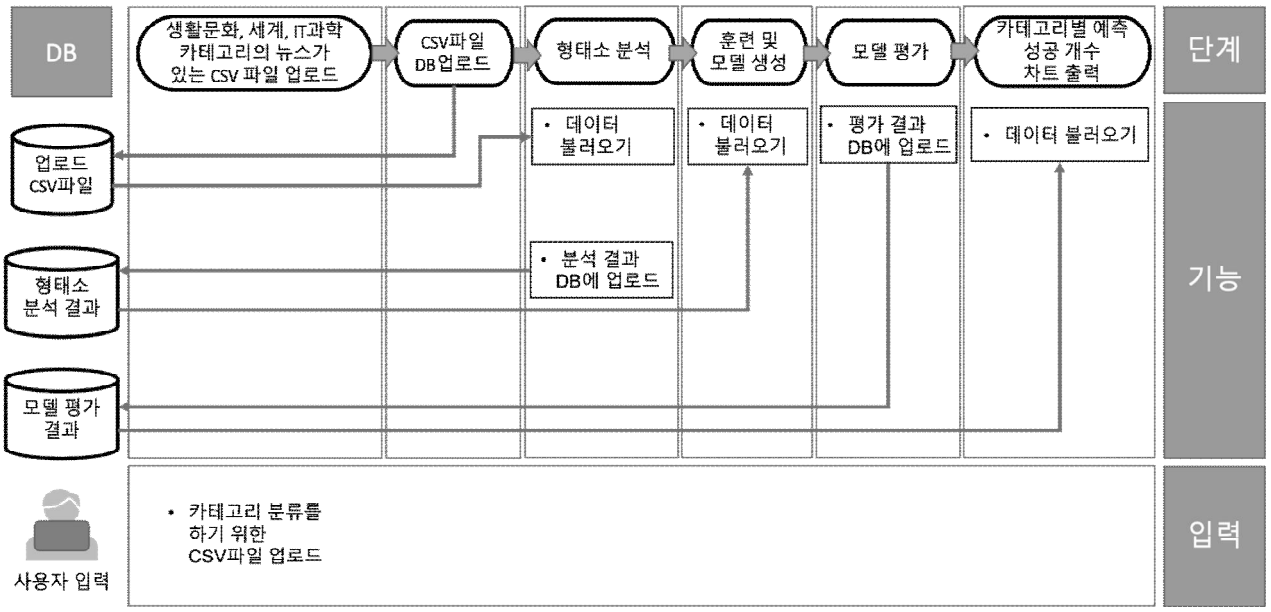

Fig. 2. System architecture

1.3 BiLSTM(Bidirectional LSTM)

BiLSTM은 텍스트 분류에서 두각을 나타내고 있는 LSTM에서 발전된 방식이다. BiLSTM은 forward LSTM과 backward LSTM을 조합해 사용한다. BiLSTM은 순차 데이터를 분류할 때 유용하며 순차적인 입력 값에 대해 이전 데이터와의 관계뿐만 아니라 이후 데이터와의 관계까지도 학습하게 된다. BILSTM은 과거의 데이터로 현재 입력을 분류하는 것과 미래의 데이터를 사용해 현재의 입력을 분류하는 것을 할 수 있다. 따라서 데이터를 분류할 때 과거와 미래의 데이터를 고려할 수 있게 된다[12].

BiLSTM과 다른 기법을 복합시킨 최근의 연구로는 BiLSTM모델과 CRF, MNB, LSTM을 사용하여 질의의 의도를 파악해 적절한 항공권을 반환하는 항공권 안내 챗봇을 구현한 연구가 있다[13]. 또한 자연어 추론에서의 교차검증 앙상블 기법으로 MRPC, RTE 데이터셋과 BiLSTM, CNN, ELMo, BERT 모델을 이용하여 기존 앙상블 기법보다 향상된 성능을 보여주는 연구가 있다[14].

III. A News Text Classification Using LSTM

본 연구는 매년 방대한 양의 콘텐츠들이 넘쳐나는 상황에서 인공지능을 적용한 카테고리 분류로 원하는 데이터를 추출하며 텍스트 분류에서 두각을 나타내고 있는 LSTM을 사용한 모델을 제안한다.

1. System architecture

본 시스템은 예시 데이터로 IT과학, 생활문화, 세계 카테고리에 속하는 한국어로 된 뉴스 텍스트 데이터를 사용하며 텍스트를 분석해 해당 3가지의 카테고리로 분류하는 모델을 생성한다. Fig. 2.은 시스템 아키텍처로 본 시스템에 대한 간결한 정보를 나타낸다.

먼저 사용자가 형식에 맞는 CSV파일을 웹 페이지에 업로드 하면 해당 CSV파일을 DB에 업로드 한 후 형태소 분석을 진행한다. 형태소 분석이 완료되면 훈련 및 모델을 생성하고 모델을 평가해 정확도를 나타낸다. 마지막으로 생성된 모델을 사용해 테스트 데이터를 예측하고 모델에 따른 예측에 성공한 비율에 대한 차트를 시각적으로 제공해준다.

2. Process

2.1 Data Preprocessing

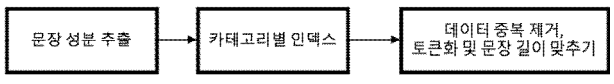

Fig. 3. Data Preprocessing

텍스트 데이터와 같은 비정형 데이터는 분석을 위해 계량화하는 과정이 필요한데 이것을 데이터 전처리라고 한다. 전처리 과정은 텍스트를 분석에 적합한 형태로 정제하는 과정을 의미한다[15]. Fig. 3.은 데이터 전처리 과정을 나타내며 아래에 그 과정에 대해 설명한다.

훈련에 사용된 데이터는 네이버 뉴스를 크롤링한 데이터이다. 카테고리는 IT과학, 생활문화, 세계이고 각각 약 7000개의 데이터를 수집했다. 수집한 데이터는 훈련하기 전 형태소 분석을 진행하게 된다. 형태소 분석이란 문장에서 최소 의미 단위인 형태소를 추출하는 과정이다[16]. 이때 의미는 어휘적 의미와 문법적 의미를 포함한다. 형태소 분석은 Mecab을 사용했고 일반명사, 고유명사, 형용사 등 총 8가지의 성분을 추출하였다.

성분을 추출한 후 생활문화 카테고리는 0으로 세계 카테고리는 1로 IT과학 카테고리는 2로 인덱스 번호를 매겨주고 뉴스 텍스트는 중복을 모두 제거한다. 만일 데이터가 중복이 되어 훈련데이터와 테스트 데이터가 중복되는 경우가 발생해 모델의 정확도를 측정하는 과정에서 오류가 생기는 것을 방지하기 위해 중복을 제거한다. 그 후 마지막 과정으로 문장의 특징을 구별하기 위해 데이터에서 상위빈도 5000개 단어만을 추출해 토큰으로 만들고 텍스트를 시퀀스로 바꿔준다. 텍스트를 시퀀스로 바꾸어주면 문장을 숫자로 나타낼 수 있게 된다. 모델의 입력으로 사용하려면 고정된 길이로 만들어야 하므로 시퀀스로 바꾼 텍스트는 최대길이를 500으로 주어 시퀀스의 길이를 일정한 길이로 맞춰주어 전처리를 완료한다.

전처리를 맞춘 텍스트 데이터는 훈련데이터와 테스트 데이터를 각각 임의로 80%, 20%의 비율로 나누어준다.

2.2 Activation function and loss function

본 연구에서 제안하는 모델은 LSTM을 활용한 학습 모델이다. 다중분류를 위해 활성화 함수로는 softmax를 사용하였고 손실함수로는 categorical cross-entropy를 사용하였다.

$$y_j = \frac{e^{z_j}}{\sum_{s=1}^{c} e^{z_s}} \quad \cdots\cdots\cdots(1)$$

식 (1)[1]은 softmax 함수를 이용한 출력이다. softmax 함수를 사용하면 각 클래스의 출력이 0과 1사이로 정규화된 값을 확률 값으로 간주하여 크로스 엔트로피 함수에 적용할 수 있다[1]. 여기서 j는 각 클래스를 의미한다. c차원의 벡터에서 j번째 원소를 z_j라 하고 j번째 클래스가 정답일 확률이 함수에 의해 도출된 y_j이라고 했을 때 softmax의 수식이며 모든 j에 대한 softmax값을 더하면 1이 나온다. 본 연구에서는 카테고리의 개수가 3개이므로 3차원의 벡터가 된다.

$$CE = -\sum_{i=1}^{N}\sum_{k=1}^{class} y_{ki}\ln\hat{y}_{ki} \quad \cdots\cdots\cdots(2)$$

식 (2)[2]는 크로스 엔트로피 손실함수이다. N은 훈련 샘플의 개수, class는 클래스의 개수이며 y_{ki}는 실제 출력을 $\hat{y}_{ki}$는 모델의 출력을 나타낸다. 손실함수에 의해 도출되는 손실 값은 실제 값과 예측 값의 오차로 신경망이 학습을 하는데 지표가 된다. 본 연구에서는 다중분류를 하고 있어 크로스 엔트로피로 categorial cross-entropy를 사용한다.

3. Learning stabilization

3.1 Dropout

본 연구에서는 정확도를 높이기 위해 LSTM을 활용한 모델에 Dropout Layer를 두 개 쌓았다. Dropout이란 Overfitting을 방지하기 위한 방법 중 하나로 Overfitting은 학습 데이터에 지나치게 집중함으로써 테스트 데이터에서의 정확성은 떨어지는 현상을 말한다. Overfitting이 발생하면 학습데이터에 대한 정확도는 향상되는 반면 테스트 데이터를 이용한 결과는 개선되지 않는 현상이 발생한다. 실제 분석 시 LSTM을 활용해 Dropout층을 추가하였을 때와 추가하지 않았을 때의 정확도는 약 2%의 차이가 나는 것을 볼 수 있었다.

3.2 GlobalMaxPool1D

GlobalMaxPool1D는 문장을 훑어가면서 나온 여러 개의 특징 벡터 정보 중 가장 큰 벡터를 골라서 반환하는 Layer이다. 즉, 문맥을 보면서 주요 특징을 골라내고 그 중 가장 두드러지는 특징을 고르는 것이다. 문장의 카테고리 분류를 하기 위해선 문장의 특징을 파악하는 것이 중요하므로 본 연구에서는 GlobalMaxPool1D Layer를 사용했을 때와 사용하지 않았을 때의 정확도에 약 2%의 차이가 있는 것을 볼 수 있었다.

4. Summary of a LSTM model

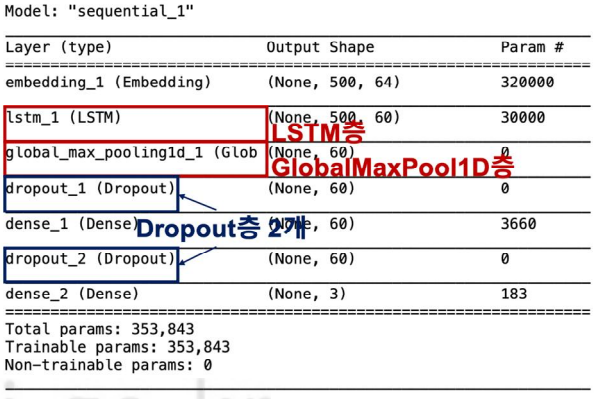

Fig. 4. Summary of a LSTM model

Table 1. RNN, LSTM, and BiLSTM model evaluation

	Model Evaluation						
Model	Val_loss	Val_acc	Train_loss	Train_acc	Epoch	Training Time	Accuracy
RNN	0.2529	0.9131	0.1147	0.9637	5	0:01:01	90.24
LSTM	0.2419	0.9266	0.1184	0.9675	5	0:02:56	92.99
Bidirectional LSTM	0.2430	0.9168	0.1121	0.9681	5	0:04:44	92.15

Fig. 4.은 LSTM을 사용한 모델의 구조이다. 가장 처음 Layer는 embedding을 하는 Layer로 언어의 벡터화 과정을 거치는 과정이다. 자연어는 수치화되지 않은 데이터이기 때문에 딥러닝에 바로 사용될 수 없다. 따라서 딥러닝에 사용할 수 있도록 embedding layer를 제일 처음에 쌓았다. 그 후 차례대로 앞서 설명한 LSTM, GlobalMaxPool1D, Dropout층을 쌓아 모델을 구성하였다.

IV. Experiment

실험을 통해 본 연구에서 제안하는 LSTM모델과 RNN, BiLSTM간의 장점과 단점을 비교해보고자 한다. 데이터는 2019년 1월 네이버의 IT과학, 생활문화, 세계 카테고리의 뉴스 데이터를 사용했다.

1. Comparing validation loss

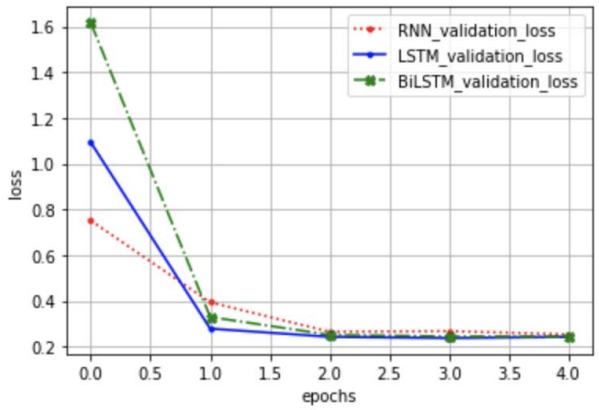

Fig. 5. Validation loss of RNN, LSTM, and BiLSTM

Fig. 5.은 RNN과 LSTM, BiLSTM을 사용한 모델의 검증 손실 값에 대한 그래프를 나타낸다. 훈련에 사용된 모델의 구조는 앞서 Fig. 4.에서 나타낸 LSTM의 모델 구성에서 LSTM의 층을 각각 RNN과 BiLSTM으로 바꾸어 구성하였다.

훈련 시 세 가지 모델 모두 5번째 훈련 이후부터 정확도가 증가하지 않고 검증 손실 값이 감소하지 않아 epoch를 5로 주어 훈련하였다. 손실 값이란 학습을 통해 얻은 추정치가 실제 데이터와의 차이를 의미하므로 0에 가까울수록 좋은 모델이라는 것을 의미한다. 세 가지 모델 모두 검증 손실 값이 점차 감소하는 것을 볼 수 있다. Table 1.를 통해 정확한 수치로 마지막 검증 손실 값이 낮은 순서대로 보자면 LSTM, BiLSTM, RNN 순으로 차례대로 0.2419, 0.2430, 0.2529이다. 따라서 LSTM의 검증 손실 값이 다른 두 모델보다 낮은 것을 알 수 있다.

2. Comparing training loss

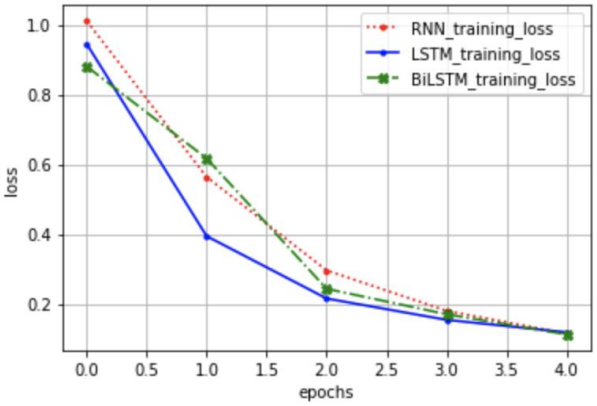

Fig. 6. Training loss of RNN, LSTM, and BiLSTM

Fig. 6.은 RNN과 LSTM, BiLSTM을 사용한 모델의 훈련 손실 값에 대한 그래프를 나타낸다.

훈련 손실 값은 계속 감소하는 반면 Fig. 5.에서 본 검증 손실 값은 3번째 이후의 훈련부터 증가하는 것을 보아 Overfitting이 일어난 것을 알 수 있다. Table 1.를 통해 정확한 수치로 검증 손실 값과 훈련 손실 값의 차이를 비교했을 때 RNN에서의 차이가 가장 큰 것으로 보아 다른 두 모델보다 RNN에서의 Overfitting이 크게 일어난 것을 알 수 있다.

3. Comparing training time

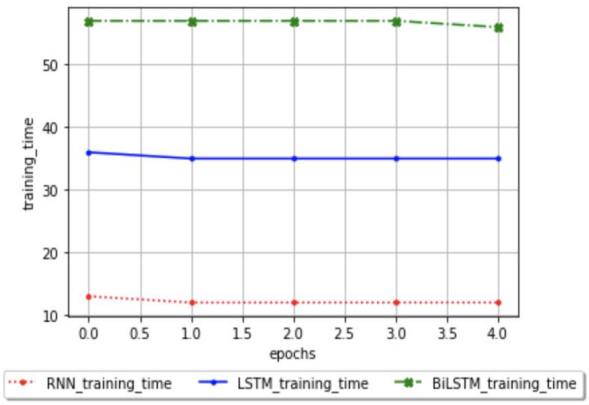

Fig. 7. Training time of RNN, LSTM, and BiLSTM

Fig. 7.은 RNN, LSTM, BiLSTM의 훈련 시간을 나타내는 그래프이다. RNN의 훈련 시간이 가장 적게 걸리는 것을 알 수 있고 LSTM은 RNN보다 약 2배가량 오래 걸리며 BiLSTM은 LSTM보다 약 2배정도 오래 걸리는 것을 볼 수 있다. 세 모델 중에서는 RNN의 훈련시간이 가장 짧지만 정확도는 가장 낮은 것을 볼 수 있었다.

4. Comparison of RNN, LSTM and BiLSTM

RNN, LSTM, BiLSTM 모두 5번째 훈련 이후부터 정확도가 증가하지 않고 검증 손실 값이 감소하지 않아 훈련 횟수를 5로 주었다. 또한 손실함수로는 동일하게 크로스 엔트로피 함수를 사용해 훈련을 하였다.

세 가지 모델의 성능을 비교하기 위한 척도로 Accuracy, Precision, Recall을 측정하였다[3]. Accuracy는 전체 예측에 대해 정답을 맞힌 비율이고 Precision은 True라고 예측한 것 중에서 실제로 True인 것에 대한 비율이다. Recall은 실제 True인 것들 중에서 True라고 예측한 것에 대한 비율로 Accuracy, Precision, Recall 모두 값이 클수록 좋은 모델로 평가된다. Precision과 Recall을 구하는 식(3)[3]은 다음과 같다.

$$precision = \frac{TP}{TP + FP} \quad \cdots\cdots\cdots(3)$$

$$recall = \frac{TP}{TP + FN}$$

한 텍스트가 A라는 카테고리에 해당된다고 가정할 때 TP, TN, FP, FN은 다음과 같이 정의된다[3].

- TP(true positive): 실제 A카테고리를 A카테고리로 예측
- TN(true negative): 실제 A카테고리가 아닌 것을 A카테고리가 아닌 것으로 예측
- FP(false positive): 실제 A카테고리가 아닌 것을 A카테고리로 예측
- FN(false negative): 실제 A카테고리를 A카테고리가 아닌 것으로 예측

Table 2. Comparison with 3 models

	Accuracy (%)	Precision	Recall
RNN	90.24	0.90	0.90
LSTM	92.99	0.93	0.92
BiLSTM	92.15	0.92	0.92

Table 2.를 통해 알 수 있듯이 LSTM을 사용한 모델이 다른 두 모델보다 우수한 것을 알 수 있다.

LSTM이 RNN의 장기의존성 문제를 해결한 인공 신경망인 것으로 보아 LSTM의 학습 능력이 더 우수한 것으로 보인다. 또한 LSTM과 BiLSTM은 Accuracy, Precision, Recall에서 큰 성능차이를 보이지 않지만 Accuracy, Precision이 LSTM에서의 측정값이 더 높은 것을 볼 수 있다.

따라서 본 연구에서는 Accuracy, Precision, Recall을 비교할 때 LSTM을 사용한 모델이 다른 두 모델보다 우수한 것을 알 수 있어 LSTM을 사용한 텍스트 분류를 권장한다.

5. A chart offered to users

사용자에게 제공되는 웹 페이지는 Spring으로 구현되었다.

사용자가 올린 파일은 RNN, LSTM, BiLSTM 세 가지를 사용해 훈련한다. 훈련 후 제공되는 차트는 훈련 후 생성된 모델을 사용해 테스트 데이터를 예측하고 예측에 성공한 비율에 대한 차트이다.

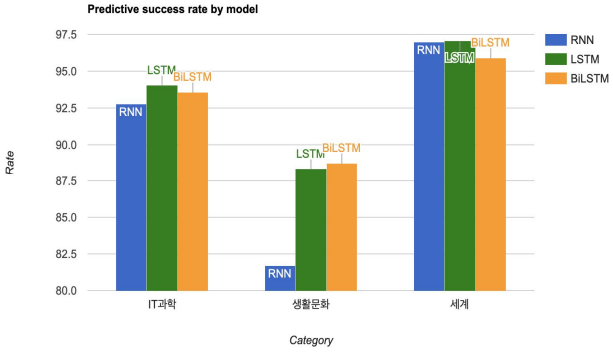

Fig. 8. Rate of successful predictions by model

Table 3. Number of successful predictions

	IT과학	생활문화	세계	총계
RNN	1,198	1,038	1,086	3,322
LSTM	1,214	1,122	1,087	3,423
BiLSTM	1,208	1,127	1,074	3,409
전체 데이터	1,291	1,270	1,120	3,681

Fig. 8.은 사용자에게 제공되는 차트이다. 바 그래프는 순서대로 RNN, LSTM, BiLSTM을 사용해 훈련하고 생성된 모델로 테스트 데이터의 카테고리를 예측하고 예측에 성공한 데이터의 비율(%)을 나타내고 있다.

Table 3.으로 볼 때 LSTM과 BiLSTM의 성능에서 큰 차이가 없고 Recall값은 동일한 것을 보았을 때 생활문화 카테고리에서는 BiLSTM이 우수한 것으로 보이지만 약 5개 차이이고 여러 번의 동일한 조건하에서의 실험마다 결과에 약간의 차이는 있더라도 정확도측면에서 생활문화부분은 유사한 것을 볼 수 있었으며 IT과학과 세계 카테고리에서의 정확도는 LSTM이 가장 우수한 것을 볼 수 있었다. 또한 비교적 LSTM과 BiLSTM의 예측 성공 개수는 비슷하지만 생활문화 카테고리의 RNN의 예측 성공 개수는 현저히 낮은 것을 볼 수 있었다.

V. Conclusions

본 연구는 매년 방대한 양의 콘텐츠들이 넘쳐나는 상황에서 모든 데이터를 개인이 분류해 필요한 정보를 찾아내기보다는 인공지능을 사용한 카테고리 분류로 원하는 데이터를 추출함으로써 편의성을 제공한다는 관점에서 시작되었다.

카테고리 분류에 적합한 모델을 찾기 위해 RNN, LSTM, BiLSTM을 활용해 3가지 모델을 사용하여 비교를 하였고 가장 적합한 모델로 LSTM을 활용한 모델을 제시한다.

현재는 뉴스 데이터를 기반으로 텍스트 분류를 진행하였지만 카테고리 분류가 필요한 또 다른 상황에서도 활용될 수 있을 것으로 기대된다.

본 연구에서는 LSTM 방식을 사용했을 때 epoch 3번째 이후부터 약간의 Overfitting이 발생하는 것을 볼 수 있었다. 따라서 이 Overfitting을 해결하면 정확도를 더 높일 수 있을 것으로 기대된다. 또한 Mecab을 사용해 형태소 분석을 하면 사용자 사전을 만들 수 있게 되는데 사용자 사전에 형태소 분석이 되지 않는 주요 단어인 AI와 같은 단어를 추가해 훈련을 한다면 정확도가 더 올라갈 것으로 예상된다.

ACKNOWLEDGEMENT

This (work) was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MOE). (No.2019전문-22)

REFERENCES

[1] Pil-Han Jeon, Sung-Kwun Oh, “Design of Robust Face Recognition System to Pose Variation Using RBFNNs-based Softmax Pattern Classifier”, Journal of Korean Institute of Intelligent System, Vol. 27, No. 6, pp. 486-492, December 2017. DOI : 10.5391/JKIIS.2017.27.6.486

[2] Sang-Beom Park, Sung-Kwun Oh, Hyun-Ki Kim, “Design of Softmax Function-based RBFNN Classifier Realized with the Aid of Optimizer Fuzzy Transform”, Journal of Korean Institute of Intelligent Systems, Vol. 28, No. 2, pp. 99-106, April 2018. DOI : 10.5391/JKIIS.2018.28.2.99

[3] Yunseok Rhee, “Malicious Code Detection Method Using LSTM Learning on the File Access Behavior”, The Journal of Korean Institute of Information Technology, Vol. 18, No. 2, pp.25-32, February 2020. DOI : 10.14801/jkiit.2020.18.2.25

[4] Artificial intelligence translator, https://www.yna.co.kr/view/AKR20210401156500054?input=1195m

[5] Face ID, http://www.bloter.net/archives/294880

[6] Identifying of cervical cancer with AI, https://www.nih.gov/news-events/news-releases/ai-approach-outperformed-human-experts-identifying-cervical-precancer

[7] Entertainment articles in Naver's'Life·Culture' section, http://www.journalist.or.kr/news/article.html?no=47749

[8] S. Jelena, M. Nijole, and M. Algirdas, "High-low Strategy of Portfolio Composition using Evolino RNN Ensembles", Inzinerine Ekonomika- Engineering Economics, Vol. 28, No.2, pp. 162-169, April 2017. DOI:10.5755/j01.ee.28.2.15852

[9] RNN and LSTM, http://colah.github.io/posts/2015-08-Understanding-LSTMs

[10] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning Precise Timing with LSTM Recurrent Networks", Journal of Machine Learning Research, Vol. 3, No. 1, pp. 115-143, January 2002. DOI:10.1162/153244303768966139

[11] Kim, Na-Rang, Adyan Marendra Ramadhani, “Automatic Classification of Civil Complaint Data Using CNN and Bidirectional LSTM : The Case of Busan, South Korea”, Korean Computers and Accounting Review, Vol. 17, No. 2, pp.81-98, December 2019. DOI: 10.32956/kaoca.2019.17.2.81

[12] Hyun-Joon Nam, Hyeon-Kyu Noh, Byung-Sub Kim, Jae-Yoon

Sim, and Hong-June Park, “On-Device Automatic Speech Recognition Application for Android Smartphone using Tensorflow Mobile Library and Bidirectional LSTM”, The Institute of Electronics and Information Engineers, (), pp. 708-711, November 2018.

[13] Yoonjae Lee, Hyeryoung Cho, Sujin Kim, Hyunseok Park, “AirScope: Implementation of Air Travel Chatbot with Named Entity Recognition by BiLSTM-CRF and Intent Analyzer by MNB, LSTM”, The Korean Institute of Information Scientists and Engineers, (), pp. 1507-1509, December 2019.

[14] Kisu Yang, Taesun Whang, Dongsuk Oh, Chanjun Park, Heuiseok Lim, “Cross-Validated Ensemble Methods in Natural Language Inference”, The Korean Institute of Information Scientists and Engineers, Vol. 48, No. 2, pp. 154-159, February 2021. DOI : 10.5626/JOK.2021.48.2.154

[15] Sae-Mi Lee, Seung-Eui Ryu, Soonjae Ahn, “Mass Media and Social Media Agenda Analysis Using Text Mining : focused on ‘5-day Rotation Mask Distribution’”, JOURNAL OF THE KOREA CONTENTS ASSOCIATION, Vol. 20, No. 6, pp.460-469, June 2020. DOI: 10.5392/JKCA.2020.20.06.460

[16] Seung-Man Lee, Young-Hun Jang, and Jung-Hwan Lim, “Implementation of a Harmful Website's Automatic Classification System based on Morphological Analysis and Skin-Color Distribution's Human Detection Algorithm”, The Korean Institute of Information Scientists and Engineers, Vol. 31, No. 1B, pp.601-603, April 2004.

Authors

Young-Dan Noh received the A.S. degree in Computer Information Engineering from Inha Technical College, Korea, 2021 respectively. Ms. Noh entered the Inha Technical College in 2018 and graduated in 2021. She has experience in IT practice and interested in IT and the fourth industrial revolution.

Kyu-Cheol Cho received the B.S., M.S. and Ph.D. degrees in Computer Science and Information Engineering from Inha University, Korea, in 2005, 2007 and 2013, respectively. Dr. Cho joined the faculty of the Department of Computer Science at Inha Technical College, Incheon, Korea, in 2016. He is currently a assistant professor in the Department of Computer Science, Inha Technical College. He is interested in cloud computing, green IT and web programming.