CNN을 활용한 교내 커뮤니티 게시글 분류 알고리즘 설계

권순욱, 남기정, 장미, 유지나*, 김재현
아주대학교 전자공학과, *아주대학교 AI융합네트워크학과
{kwonsw422, ng0716, tjwls0868, *jina1114, jkim}@ajou.ac.kr

Design of Classification Algorithm for Campus Community Posts using CNN

Soon Wook Kwon, Gi Jung Nam, Mi Jang, Jina Yu*, Jae-Hyun Kim
Department of Electrical and Computer Engineering, Ajou Univ.,
*Department of Artificial Intelligence Convergence Network, Ajou Univ.

요 약

본 논문은 convolutional neural network (CNN)을 사용하여 대학교 온라인 익명 커뮤니티인 에브리타임에 게시된 글의 카테고리 분류 및 질문성 여부를 판별하는 알고리즘을 설계한다. 크롤링을 통해 데이터를 수집하고 6개의 카테고리 및 질문성 여부를 판별하는 총 7개의 학습모델을 생성한다. 알고리즘은 데이터의 학습모델별 유사도를 출력하고, 유사도가 임계값 이상인 경우 해당 카테고리를 갖는 것으로 판단한다. 학습모델의 평균 정확도는 96.89%이며, 새로운 데이터에 대하여는 94.66%의 예측 정확도를 보인다. 따라서 설계한 알고리즘을 통해 게시글이 적절한 카테고리로 분류된다.

Ⅰ. 서론

최근 COVID-19 팬데믹 상황의 여파로 대부분의 대학교 수업이 비대면으로 진행됨에 따라 동기나 선후배 간의 교류 기회 감소로, 대학생들이 온라인 커뮤니티에서 정보 습득을 의존하는 경향을 보인다 [1]. 그러나 최대 커뮤니티인 에브리타임을 살펴보면 다양한 주제의 글이 게시되지만, 게시글의 카테고리가 분류가 되지 않으므로 원하는 정보만 골라서 볼 수 없기 때문에 정보 습득의 효율성이 낮고 편의성이 떨어지는 문제점이 있다 [2].

본 논문에서는 비대면 상황에 따른 정보 습득의 효율성과 커뮤니티 이용의 편의성 증진을 위해 CNN 모델을 통한 게시글 분류 알고리즘을 설계한다. 설계한 알고리즘은 게시글을 6개의 카테고리로 분류하고, 게시글의 질문성 여부를 판별한다. 게시글 분류 알고리즘을 통해 사용자들은 원하는 주제의 게시글만 확인 가능하므로 커뮤니티 이용의 편의성 증진을 기대할 수 있다. 질문성 여부는 의문문으로 쓰인 게시글 외에도 정보 습득이 목적인 게시글까지 포함하여, 정보 습득의 효율성 증진을 기대할 수 있다.

Ⅱ. 본론

가. 입력 데이터

1) 데이터 수집

본 논문에서는 데이터 수집을 위해 Python의 Selenium 라이브러리를 이용한다. 커뮤니티 내 자유게시판 페이지 인덱스에 존재하는 게시물의 제목과 내용 오브젝트를 크롤링하고 이를 하나의 문자열로 병합한 것을 데이터셋으로 사용한다. 데이터셋은 아주대학교 에브리타임 자유게시판에 2021년 9월 3일부터 2022년 4월 14일까지의 기간 동안 게재된 총 55,614건의 게시글의 제목 및 내용을 바탕으로 데이터화한 것이다.

2) 카테고리 선정 및 데이터 라벨링

카테고리 선정을 위하여 먼저 수집한 데이터를 KoNLPy 라이브러리의 형태소 분석기 Kkma를 사용하여, 사용 빈도수가 높은 50개의 명사 데이터를 추출한다. 명사 데이터로 정해진 카테고리 목록에서 다음과 같은 조건을 만족하는 카테고리를 분류 대상으로 최종 선정한다. (1) 총 게시글 대비 해당 카테고리 게시글 비율이 2% 이상일 것. (2) 카테고리 분류의 목적이 뚜렷할 것. (3) 월드컵 같은 특정 기간에만 언급되는 주제가 아닐 것. (4) 커뮤니티 내 갈등 및 분란을 야기하는 주제가 아닐 것. 표 1은 해당 조건을 만족하는 6개의 카테고리와 질문성 항목의 세부 판별 기준이다.

다음으로 정해진 카테고리 및 판별 기준을 바탕으로 수집한 데이터를 직접 확인하여 게시글의 카테고리 및 질문성 여부를 라벨링한다. 이 때, 각 게시물은 그 내용에 따라 여러 개의 카테고리에 동시에 포함될 수 있으며, 모든 카테고리에 포함되지 않는 경우도 존재한다. 각 카테고리 및 질문성 게시물의 건수와 총 데이터 대비 비율은 표 2와 같다.

표 1. 게시글 분류 기준안

게시글 분류	포함 게시글의 내용
수업	특정 과목, 수업관련 일반적인 내용
인간관계	타인과의 관계, 이성/매력, 성격
학과	특정학과, 단과대학 관련 내용
홍보&구인&동아리	홍보성 게시글, 동아리, 소학회
취업&진로&행정	장학, 진로관련 프로그램, 취업 관련 내용
주변장소	교내 건물, 대학교 근처 장소
질문성	정보 습득이 목적인 경우

표 2. 분류된 게시글 수 및 비율

번호	게시글 분류	게시글 수(건)	비율(%)
1번	수업	22,110	39.76
2번	인간관계	1,110	2.00
3번	학과	2,521	4.53
4번	홍보&구인&동아리	3,615	6.50
5번	취업&진로&행정	4,000	7.19
6번	주변장소	6,132	11.02
7번	질문성	38,115	68.53

나. 데이터 전처리

1) Okt를 이용한 형태소 분석 및 사전 튜닝

KoNLPy 라이브러리에 내장된 Okt 형태소 분석기를 사용하여 데이터를

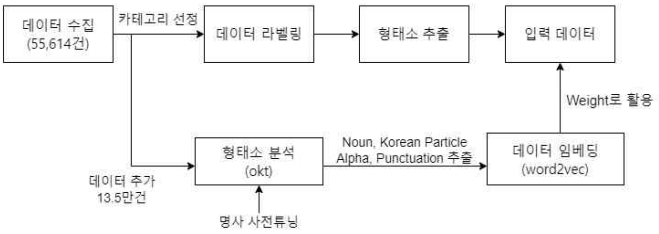

그림 1. 데이터 전처리 과정 프로세스 도식

형태소 단위로 나눈다. 다만, 커뮤니티 특성상 아주대학교의 수업명, 사람 이름으로 추정되는 자음 배열, 학교 건물 및 주변 가게들의 이름, 일반적으로 사용되는 줄임말 등을 다수 포함하고 있다. 이러한 단어들은 카테고리 분류에 중요한 역할을 하나, 형태소 분석기 사전에 내장되어있는 단어가 아니므로 잘못 분석될 가능성이 크다. 따라서 데이터 라벨링 과정에서 잘못 분석될 수 있는 형태소 목록을 수집하고, Okt 형태소 분석기 사전에 추가하는 사전 튜닝 과정을 통하여 형태소 분석의 정확성을 높인다.

2) Word2Vec 임베딩

Word2Vec은 단어를 벡터 공간에 매핑하고, 벡터 값을 계산하여 단어 간 유사도를 예측하는 시스템이다 [3]. 먼저 Word2Vec 함수에 입력으로 들어갈 단어 list를 정한다. list에 저장되는 단어들은 특정 품사의 단어들로 설정하며, 학습 데이터의 특성을 고려하여 “Noun”, “Punctuation”, “KoreanParticle”, “Alpha”, 품사들만 저장한다. “Noun”은 일반적인 명사로, 모든 카테고리 분류에 사용한다. “Punctuation”은 문장부호로, 질문성을 판별하는 학습모델에서 추가로 사용한다. “KoreanParticle”는 한국어 자음 및 모음으로, 교내 커뮤니티 내에서는 교수님의 성함 혹은 과목명으로 사용되어 수업 카테고리와 연관성이 있다. “Alpha”의 경우 영문 알파벳으로, 학점과 관련이 있다.

다음으로는 Word2Vec 라이브러리를 통하여 list에 저장된 단어들의 연관성을 파악한다. 임베딩 정확도를 높이기 위해, 라벨링한 데이터 외에도 추가 크롤링을 통해 수집한 데이터를 포함하여 총 19만 건의 데이터로 임베딩한다. 알고리즘은 skip-gram 모델을 사용하고, vector size는 200으로 설정하고, 최소 2번 이상 나온 단어에 대해서 학습하도록 하며, windows를 4로 설정하여 양쪽으로 총 4개의 단어를 고려하도록 한다.

다. 모델 학습

1) CNN 모델을 통한 이진분류

클래스를 이진분류하는 CNN 모델로써 Yoon Kim Convolutional Neural Networks for Sentence Classification 모델을 적용하여 학습을 진행한다 [4]. 각 카테고리 및 질문성 여부에 대해 학습을 진행하여 총 7가지 학습모델을 생성한다. 각 학습모델별로 이진분류에 사용되는 positive 데이터에 대한 negative 데이터의 비율을 약 100%~120% 정도로 조절하여 학습 데이터 비율을 맞춘다. 모델의 filter size는 10이며 dropout은 0.5~0.8로 설정하고 activation 함수는 rectified linear unit (ReLU), optimizer는 adaptive moment estimation (adam)을 사용한다. Weight는 Ⅱ-나-2) 과정에서 Word2Vec을 이용하여 임베딩한 데이터로 설정한다. 생성된 학습모델들은 데이터의 카테고리 유사도를 수치화하고 해당 값이 임계값 이상인 경우, 해당 카테고리를 갖는 것으로 판단한다.

2) 성능평가

성능평가를 위해 각 카테고리별 학습모델의 정확도를 판별한다. 카테고

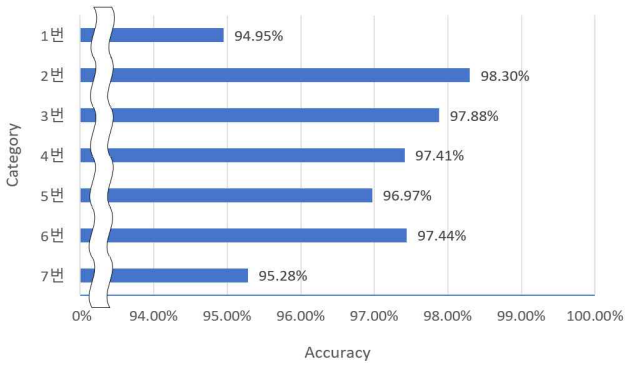

그림 2. 학습모델별 정확도

표 3. 새로운 데이터에 대한 예측 정확도

게시글 분류	정확도(%)
카테고리 평균	94.40
질문성	96.20
예측 평균 정확도	94.66

리별 정확도는 그림 2와 같으며 7개 학습모델의 평균 정확도는 96.89%이다. 새로운 게시글에 대한 사용 적합성을 판별하기 위해 추가적으로 500건의 새로운 데이터에 대해 라벨링을 하고, 각 학습모델로 카테고리 및 질문성 여부 예측을 진행한다. 표 3은 새로운 데이터에 대한 각 항목의 예측 정확도이다. 7개 학습모델의 예측 평균 정확도는 94.66%로, 학습모델이 새로운 게시글의 카테고리 및 질문성 여부를 적절하게 분류하는 것을 알 수 있다.

Ⅲ. 결론

본 논문은 대학생들의 주된 정보 습득 방식에 맞추어 커뮤니티에서의 정보 습득 효율성과 편의성 증진을 목표로 한다. 아주대학교 커뮤니티 게시글을 CNN 구조를 기반으로 한 자연어 처리 과정을 통하여 관심도 높은 카테고리들로 분류한다. 각 카테고리별 및 질문성 게시글 판별 정확도는 평균 96.89%의 정확도를 보이며, 새로 게시되는 게시글에 대하여는 평균 94.66%의 정확도를 보인다.

제시한 알고리즘을 통해 사용자가 원하는 카테고리의 정보 습득 효율성을 높임에 따라 사용자들의 커뮤니티 이용의 편의성 증진과 이로 인한 커뮤니티 이용률 증가를 기대할 수 있다. 이에 따라 교내 커뮤니티를 통해 얻는 정보의 양과 질적인 측면 둘 다 향상시킬 수 있을 것으로 기대한다.

참 고 문 헌

[1] Sinae Choe, Sanghee Oh, “Everyday Life Information Behaviors of College Students on Online Communities: A Case Study of Everytime”, The Journal of Korean Library and Information Science Society, Volume. 52, Issue 3, pp.239-266, Sep, 2021.

[2] 포쓰저널, “대학생 필수앱에 에브리타임,인스타페이,배달의민족 등 선정”, 2022, (https://www.4th.kr/news/articleView.html?idxno=2013904).

[3] 고동우, 양정진, “KoNLPy와 Word2Vec을 활용한 한국어 자연어 처리 및 분석”, 한국컴퓨터종합학술대회 논문집, pp.2140-2142, Jun, 2018.

[4] Yoon Kim, “Convolutional Neural Networks for Sentence Classification”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP), Association for Computational Linguistics, pp.1746-1751, Aug, 2014.