특집논문 (Special Paper)
방송공학회논문지 제24권 제1호, 2019년 1월 (JBE Vol. 24, No. 1, January 2019)
https://doi.org/10.5909/JBE.2019.24.1.77
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

문서 유사도를 통한 관련 문서 분류 시스템 연구

정 지 수[a], 지 민 규[a], 고 명 현[b], 김 학 동[b], 임 헌 영[b], 이 유 림[c], 김 원 일[d]‡

Related Documents Classification System by Similarity between Documents

Jisoo Jeong[a], Minkyu Jee[a], Myunghyun Go[b], Hakdong Kim[b], Heonyeong Lim[b], Yurim Lee[c], and Wonil Kim[d]‡

요 약

본 논문은 머신 러닝 기술을 이용하여 과거의 수집된 문서를 분석하고 이를 바탕으로 문서를 분류하는 방법을 제안한다. 특정 도메인과 관련된 키워드를 기반으로 데이터를 수집하고, 특수문자와 같은 불용어를 제거한다. 그리고 한글 형태소 분석기를 사용하여 수집한 문서의 각 단어에 명사, 동사, 형용사와 같은 품사를 태깅한다. 문서를 벡터로 변환하는 Doc2Vec 모델을 이용해 문서를 임베딩한다. 임베딩 모델을 통하여 문서 간 유사도를 측정하고 머신 러닝 기술을 이용하여 문서 분류기를 학습한다. 학습한 분류 모델 간 성능을 비교하였다. 실험 결과, 서포트 벡터 머신의 성능이 가장 우수했으며 F1 점수는 0.83이 도출되었다.

Abstract

This paper proposes using machine-learning technology to analyze and classify historical collected documents based on them. Data is collected based on keywords associated with a specific domain and the non-conceptuals such as special characters are removed. Then, tag each word of the document collected using a Korean-language morpheme analyzer with its nouns, verbs, and sentences. Embedded documents using Doc2Vec model that converts documents into vectors. Measure the similarity between documents through the embedded model and learn the document classifier using the machine running algorithm. The highest performance support vector machine measured 0.83 of F1-score as a result of comparing the classification model learned.

Keyword : Document analysis, Related document, Doc2Vec, Machine learning, Document classification

a) 세종대학교 소프트웨어융합학과(Department of Software Convergence, Sejong University)
b) 세종대학교 디지털콘텐츠학과(Department of Digital Contents, Sejong University)
c) 세종대학교 인공지능언어공학과(Department of Artificial Intelligence and Linguistic Engineering, Sejong University)
d) 세종대학교 소프트웨어학과(Department of Software, Sejong University)
‡ Corresponding Author : 김원일(Wonil Kim)
E-mail: wikim@sejong.ac.kr
Tel: +82-3408-3795
ORCID: https://orcid.org/0000-0002-1489-8427

※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2015R1D1A1A01060693)
· Manuscript received November 16, 2018; Revised December 27, 2018; Accepted January 10, 2019.

Ⅰ. 서 론

정보사회에서 문서의 수는 기하급수적으로 늘어나고 있다. 과거의 문서들은 현재 사건들의 근거가 되기 때문에 웹상에서 문서들을 수집하고 분석하는 일은 매우 중요하다. 이러한 작업을 하기 위해 텍스트 데이터를 다루는 일이 필요하다[1].

텍스트 마이닝(text mining)은 비구조적인 텍스트 문서로부터 정보를 찾아 지식을 발견하는 것으로 텍스트 분류는 텍스트 마이닝 연구 분야의 부분이다[2][3]. 목적에 따라 정보를 추출하기 위해서는 분석을 통해 분류를 할 필요가 있다. 문서 분류를 위해서는 텍스트 데이터를 정형 데이터로 변환해야 하고 이러한 연구에는 통계적 정보 기반, 신경망 기반 방법 등이 존재한다.

정형 데이터로 변환하고 문서를 분류하는 과정에서 다양한 방법이 제시된다. 텍스트 데이터를 정형화하는 방법으로 벡터값으로 표현하는 방법이 일반적이며, 분류의 성능을 높이기 위해서 연구 과정이 진행되고 있다. 본 연구에서는 단어를 벡터로 표현하여 단어간의 거리로 단어의 의미와 유사도를 알 수 있는 단어 임베딩에 대해 주목하였다. 단어만이 아닌 문장, 단락, 문서와 같은 길이의 텍스트를 벡터로 표현하고, 문서간의 유사도를 측정할 수 있는 문서 임베딩 모델에 주목하였고, 해당 모델을 통한 분류 과정이 미흡하다 판단하여 직접 데이터 수집을 위한 웹 크롤러를 설계하고 데이터의 일부를 분류 목적에 따른 전처리 과정을 진행하였다. 정제된 데이터를 문서 임베딩을 위한 모델로써 학습한다. 학습된 모델을 토대로 분류 알고리즘으로 가장 뛰어난 성능을 보인 4가지의 머신 러닝 기술을 사용하여 각각 성능을 비교 분석한다.

본 연구는 문서 유사도를 통한 임베딩 모델을 토대로 분류기에 주로 사용되는 4개의 머신 러닝 알고리즘을 비교 분석한다. 웹 크롤러를 통해서 정해진 카테고리내의 분류 문서들이 아닌, 수집할 특정 도메인과 관련된 키워드를 필터링하여 과거 문서들을 수집한다. 텍스트 데이터를 벡터로 표현하는 Doc2Vec 모델에 대하여 연구하고 이를 통해 수집한 문서를 임베딩한다. 그리고 머신 러닝 기술을 사용해서 문서의 주제를 분류하는 방법을 연구한다.

본 논문의 구조는 다음과 같다. 제2장에서 이전 연구를, 제3장에서 본 논문에서 제안하는 문서 분석을 통한 관련 문서 분류 시스템에 대한 설계를 소개한다. 제4장에서는 제안 모델을 이용한 실험을 진행하고 결과를 기술하였다. 제5장에서는 본 연구의 결론 및 향후 연구 방향에 대해 논의한다.

Ⅱ. 이전 연구

1. TF-IDF

자연어 처리 기술에서 비정형 텍스트 데이터를 정형화하는 것은 텍스트 데이터를 다루기 위한 가장 기본적이며 중요한 과정이다. 정형화를 위해 TF-IDF(Term Frequency-Inverse Document Frequency)와 같은 통계적 방법을 이용할수 있다. TF는 특정 단어가 문서 내에서 등장하는 빈도를 토대로 정해지는 가중치가 적용된 통계적 수치이다. IDF는 전체 문서군 크기에서 특정 단어가 나타난 문서 수를 나누어서 구할 수 있다. TF-IDF는 TF와 IDF를 곱한 값으로 해당 단어가 문서의 특징을 구분하는지 평가하며, 범용적으로 널리 사용되고 있는 통계적 기법이다. 이전 연구에서 TF-IDF를 이용하여 키워드 추출을 통해 추천 시스템에 응용하였다. 문서 분류에서 핵심적인 키워드 추출을 TF-IDF를 이용해 TF-IDF값이 높은 단어일 때 추출된 키워드가 문서 분류를 하는 것에 있어 중요하다. 추출된 키워드가 곧 문서 분류의 라벨과 연관이 있으며 이러한 TF-IDF방식은 문서 분류에 있어 좋은 결과를 나타낸다[4]. 국외의 이전 연구[5]에서 웹 문서의 검색 결과 성능을 향상시키기 위해 주제어를 추출하여 유사도에 따라 문서를 분류하였다. 문서 분류에 있어 TF-IDF는 문서의 내용을 대표하는 주제어를 추출하는데 유용하게 사용되어 왔지만 빈도 정보에 의한 방법은 한계가 있다. 따라서 본 논문은 빈도 정보뿐만 아니라 단어의 의미를 알 수 있는 임베딩 모델에 주목하였다. 문서의 내용에 대한 유사도를 측정하는 임베딩 모델을 학습한 후, 분류 모델의 비교 분석을 하여 성능향상에 기여한다.

2. Word2Vec

단어를 벡터로 표현하는 과정을 단어 임베딩(Word embedding)이라고 부르며, 단어 임베딩을 위해서는 벡터 공간 모델(Vector Space Model)이 주로 사용된다[6]. 벡터 공간 모

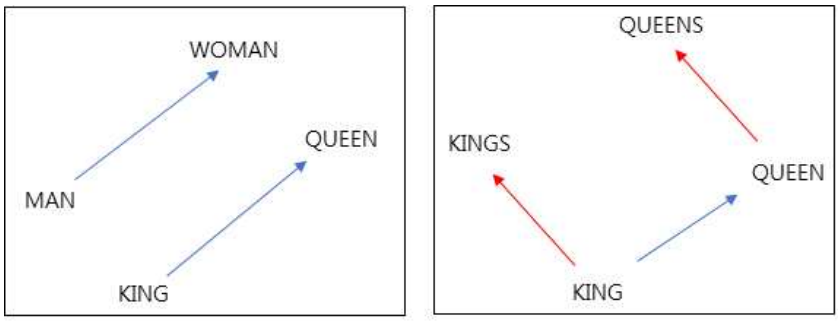


그림 1. 벡터 거리로 단어간의 의미 파악
FIg. 1. Understand the meaning between words in vector distance

델은 벡터 공간 내 주변에 등장하는 단어는 서로 비슷한 의미를 가진다는 분산 가설(Distributional Hypothesis)(Harris, 1954)에 기반을 둔다. 의미상 유사한 단어를 서로 인접시켜 임베딩 시키므로 벡터로 표현된 단어들은 문법적인 부분만 아니라 의미적인 부분까지 반영된다. 결론적으로, 단어 임베딩을 통해 생성된 벡터에 따르면 거리가 가까울수록 단어들은 서로 비슷한 벡터를 가지게 된다. 또한 단어 벡터에 포함된 단어들은 모두 수치화되어 있기 때문에 단어와 단어 간의 거리를 활용한 벡터 연산이 가능하고 이로 인해 추론을 할 수 있게 된다[7].

그림 1은 단어간의 거리를 통해 유사도를 알 수 있음을 보여준다. 벡터화하여 단어간의 거리를 계산하여 측정, 추론할수 있다면 단어간의 의미를 알 수 있고, 이러한 방법으로 정형화을 하는 것을 Word2Vec이라고 한다. Word2Vec을 통해 학습한 모델로 분류기를 만든 이전 연구[8]들이 존재하고, 단어 임베딩 모델을 통한 분류기는 좋은 성능을 보인다. 단어만을 벡터로 표현하는 것이 아닌 문장, 단락, 문서와 같은 가변 길이의 텍스트를 벡터로 표현하기 위해 확장형인 Doc2Vec 모델이 제안되었고, 해당 방식은 문서 분류, 감정 분석, 정보검색에 좋은 성능을 보인다[9].

본 연구에서는 데이터 수집을 위해 관련 키워드를 필터링하여 특정 도메인에 관한 문서를 수집하기 위한 웹 크롤러를 만든다. 수집한 일부 데이터를 학습데이터로 사용한다. 워드 임베딩 방식의 하나인 Doc2Vec 모델의 품질을 높이기 위해 전처리 과정을 진행한 후, 모델을 학습한다. 기존 분류기 머신 러닝 기술을 사용하여 분류 모델 평가를 나타낸다. 실험한 결과로 비교 분석하고, 개선사항 및 향후 연구 방향에 대해 논의한다.

Ⅲ. 제안 방법

본 연구에서 제안하는 문서 분류 시스템의 전체 구조는 그림 2와 같다. 웹 크롤러를 이용하여 특정 도메인에 관한

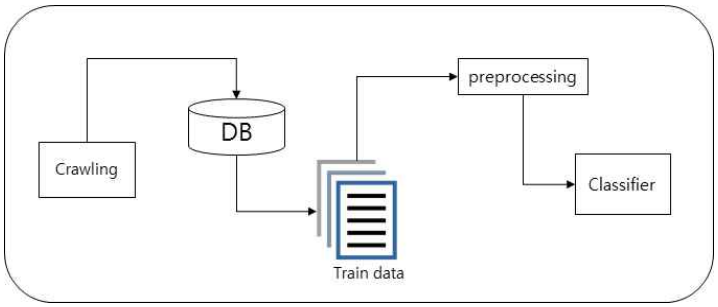


그림 2. 문서 분류기의 전체적인 설계
FIg. 2. Overall Design of Document Classification

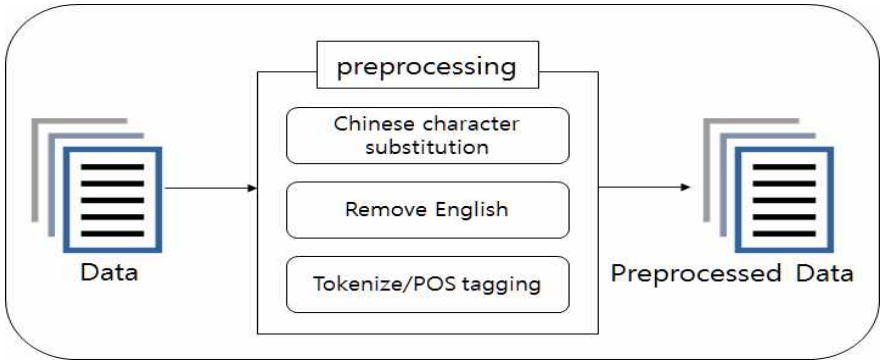


그림 3. 문서 분류기의 데이터 전처리 과정
Fig. 3. Document calssification data preprocessing

문서를 수집하고 데이터베이스에 저장한다. 그림3은 저장한 문서에 대하여 전처리를 진행하고 문서를 벡터로 임베딩하는 Doc2Vec모델을 학습한다. 그리고 머신 러닝 기술을 이용하여 문서 분류기를 학습한다.

1. 웹 크롤러 설계

문서 분류기에 사용할 학습용 데이터를 수집하기 위해 웹 크롤러를 설계한다. 현재 인터넷에는 매우 방대한 양의 문서 데이터들이 존재하므로 모든 문서 데이터를 수집하기 어렵다. 따라서 분류에 필요한 데이터를 가져옴과 동시에 데이터의 양을 줄이기 위해 특정 도메인을 선정하고 해당 도메인에 대한 문서를 수집한다. 특정 도메인과 관련된 뉴스 기사를 선정하고, 관련된 키워드(Keyword)를 중심으로 포함 여부에 따라 필터링하는 방식으로 웹 크롤러를 설계하였다.

그림 4의 웹 크롤러는 특정 도메인에 관한 문서를 수집하기 위해 키워드를 기반으로 필터링 작업을 거친다. 수집한

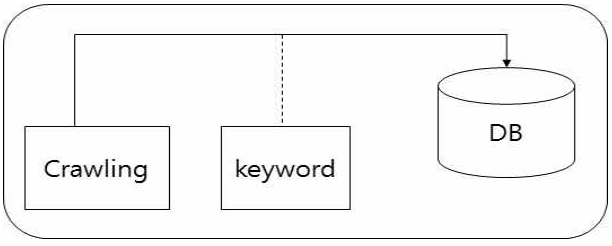


그림 4. 키워드를 기반으로 수집하는 웹 크롤러 설계
Fig. 4. Web crawler designs collected based on keywords

데이터는 HTML 태그를 제거하고, 필요 부분만 추출하는 전처리 과정을 거쳐 뉴스의 제목(Title)과 뉴스의 본문(Content)을 데이터베이스에 저장한다.

2. 데이터 전처리

데이터 전처리 과정은 데이터가 결측치(none value), 이상 값(outlier), 에러(error), 잡음(noise) 등을 포함하여 발생하는 문제를 해결하기 위한 필수적인 과정이며 데이터 전처리를 통해 모델의 성능 향상에 기여한다.

본 연구에서는 데이터 전처리 과정을 통해 특수문자와 같은 불필요한 문자들을 제거하고, 오래된 뉴스인 경우 본문 내용에 한자가 있어 한자를 음독으로 치환한다. 기사 내용의 불필요한 기자 이메일 주소와 영어 단어 및 문장을 제거한다. 다음으로 한글 형태소 분석 및 품사 태깅을 지원하는 라이브러리 KoNLPy의 Komoran클래스를 사용하여 형태소 분석을 실시한다[10]. 형태소 분석을 통해 동일한 단어라도 품사를 결합하면 의미를 구분 할 수 있기 때문에 모델의 성능 향상에 기여한다.

3. Doc2vec 모델 학습

텍스트 데이터를 머신 러닝 기술에 적용하기 위해서 문서를 벡터 값으로 변환해야 한다. 단어의 순서와 의미는 중요한 정보로, 이를 표현하기 위해 단어를 벡터화하는 Word2Vec 모델의 확장형인 Doc2Vec 모델을 사용한다. 단

어만이 아닌 단어들로 이루어진 문장, 단락, 문서와 같은 가변 길이의 텍스트를 벡터로 표현하기 위한 방법으로 Doc2Vec이 제안되었다. Doc2Vec은 문장, 단락, 문서와 같은 길이의 텍스트를 임베딩하는 모델로써, Word2Vec 알고리즘을 문장, 단락 또는 전체 문서와 같이 더 큰 텍스트 블록에 대한 연속 표현을 비지도학습 하도록 수정된 모델이다. Doc2Vec은 모델을 훈련할 때 라벨과 실제 데이터가 필요한데, 이 때의 라벨은 문서의 주제, 빠진 내용이 사용되거나 일반적으로 문서의 파일명, 문서 번호, 혹은 본 연구에서 사용할 2가지의 문서 분류 라벨명으로 사용된다. 문서간의 유사도를 측정하여 문서 분류의 성능을 높이고 라벨의 정의된 데이터를 정확히 분류하여 출력 결과를 보인다. Doc2Vec은 문서 분류, 감정 분석, 정보 검색에 적합하나, 실질적으로 문서 분류에 사용한 경우가 TF-IDF에 비해 적어 본 연구에서 제안하는 Doc2Vec 모델을 이용하여 문서를 임베딩하는 모델을 학습한 후, 관련 문서를 분류하고 성능을 평가한다. 문서 단어의 빈도수를 가중치로 사용한 TF-IDF방식과는 달리 문서를 벡터로 표현하고 유사도를 통해 문서 분류하는 방식에 주목하였다.

그림 5는 Doc2Vec 모델 학습을 위한 방법으로 DBOW (Distributed Bag Of Words) 방식과 PV-DM(Distributed Memory Model of Paragraph Vector) 방식으로 2가지 학습 방식 중, 문서 벡터와 단어 벡터 정보를 모두 사용하는 PV-DM 모델 방식을 이용하였다[11]. 그림 5는 PV-DM 모델의 구조를 나타낸다. d는 문서를, w는 단어를 의미한다. N은 벡터의 크기를 나타내고, V는 전체 단어의 수를 의미한다.

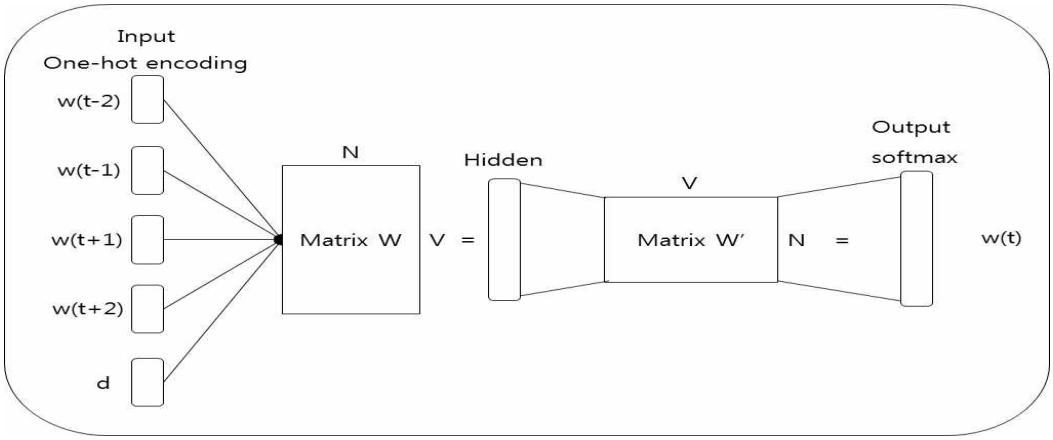


그림 5. Doc2Vec 모델 구조 (PV-DM)
Fig. 5. Doc2Vec Model structure (PV-DM)

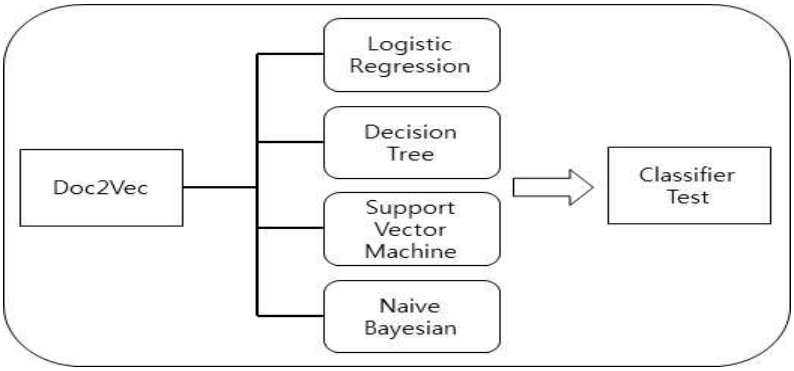


그림 6. 머신 러닝 기술 분류기 실험
Fig. 6. Machine Learning Algorithm Classifier Experiment

Input에서는 단어를 one-hot encoding으로 넣어주고, 주변의 단어 w(t-2, t-1,t+1, t+2)를 각각 projection 시킨 후 그 벡터들의 평균을 구해서 Hidden에 보낸다. 다음으로 여기에 Weight Matrix를 곱해서 Output으로 보내고 softmax 계산을 한다. 계산한 결과를 타깃 단어인 w(t)의 one-hot encoding과 비교하여 에러를 계산한다. 이때의 d는 메모리와 같은 기능으로 현재의 문맥에서 빠진 것이나 단락에서의 주제를 기억해주는 역할을 한다. 본 연구에서 d는 특정 주제로 분류된 문서들과 그 외의 문서 범주를 의미한다.

4. 분류기 설계

Doc2Vec 모델을 학습한 후, 학습된 모델에서 추출한 문서 벡터로 각각의 뉴스 데이터들을 관련 문서로 분류할 수 있도록 4개의 머신 러닝 기술을 사용한다. 각각의 머신 러닝 기술은 분류 모델에 있어 좋은 결과를 나타내는 머신 러닝 기술로, 학습된 Doc2Vec 모델을 토대로 머신 러닝 기술을 적용하여 분류 모델의 실험에 사용하였다.

그림 6은 학습된 Doc2Vec 모델을 토대로 사용한 머신 러닝 기술의 실험 과정을 나타낸다. 첫 번째로 반응 변수가 1 또는 0인 이진형 변수에서 쓰이는 분류 방법의 일종인 로지스틱 회귀분석(Logistic Regression)[12]이다. 두 번째로 데이터를 분석하여 이들 사이에 존재하는 패턴을 예측 가능한 규칙들의 조합으로 나타내는 방법인 결정 트리(Decision Tree)[13], 세 번째로 주어진 데이터 점들이 두 개의 클래스 안에 각각 속해 있다고 가정했을 때, 새로운 데이터 점이 두 클래스 중 어느 곳에 속하는지 결정하는 것이 목표로 하는 분류 방법인 서포트 벡터 머신(Support Vector Machine)[14]이다. 마지막으로 사건 B가 발생한 경우 A의 확률을 나타내는 베이스 정리(Bayesian probability)를 기반으로 하는 분류 방법인 나이브 베이스(Naive Bayesian)[15]이다. 이 4개의 머신 러닝 기술을 사용한 분류 모델을 비교 분석한다. 본 연구에서는 특정 도메인의 뉴스 기사를 분류하고, 분류의 목적으로 뉴스 기사의 내용이 위해한 주제의 뉴스인지, 위해하지 않은 지에 대한 뉴스를 분류하기 위해 2가지의 라벨을 통해 학습 데이터를 구축한 후 모델을 비교 평가한다.

Ⅳ. 실험 결과 및 논의

1. 실험 데이터

학습데이터를 만들기 위해 본 연구에서는 연구에 사용하기 위한 데이터 수집을 할 필요가 있다. 실험에 사용하기 위한 데이터를 공개적으로 배포되지 않으므로 특정 도메인에 관한 데이터를 수집하기 위한 웹 크롤러를 통해 진행하였다. 연구 주제인 “생활화학제품”에 관련된 품목과 위해관련 키워드를 필터링과정을 통해 뉴스 기사를 수집하였다. 수집한 4만 건의 데이터 중 6000개의 데이터를 학습 데이터로써 사용하였다. 그 후 수작업을 통해 라벨링 과정을 진행하여 위해도 관련 뉴스와 관련 없는 뉴스 데이터로 분류하였다.

표 1에서 News class에서 Event news는 위해도 관련 뉴스를 의미하고, 해당 라벨링은 ‘1’으로 지정하고, content에서는 라벨링 된 뉴스 데이터의 내용을 의미한다. Other news는 그 외의 관련되지 않은 뉴스를 나타내고 해당 라벨링은 ‘0’으로 표기하고, content에서는 라벨링 된 해당 뉴스의 내용을 의미한다. 실험을 위하여 4200개를 훈련용 데이

표 1. 위해도 관련 뉴스와 그 외의 뉴스의 라벨링된 데이터 예시
Table 1. Examples of hazad related News and other news labeled data

News class	number	content
Event news	1	반영구 화장용 문신 염료에서 발암물질과 중금속이 다량 검출돼 주의가 필요하다. 유해 물질 중 니켈은 피부 알레르기를 잘 유발시키는 대표적인 금속물질이다. (Because a large amount of carcinogens and heavy metals are detected in the dye for cosmetics of semitransparent cells, care needs to be taken. Nickel is a representative metal substance that causes skin allergies well.)
Other news	0	그린을 팔지 않으면 장사가 안된다는 환경시대를 맞아 녹황색 채소 수세미 은행잎 등 식물을 원료로 쓴 자연화장품이 잇따라 선보이고 있다. (In the wake of the environmental era when green is not sold, natural cosmetics such as green vegetables, susemi, and ginkgo leaves are showing off one after another.)

터로, 1800개를 검증용 데이터로 분류하였다. 본 연구에서는 "생활화학제품"에 대해 위해한 뉴스와 그렇지 않은 뉴스로 분류하기 위해 2개의 라벨링과정을 진행하였다.

2. 실험 과정

실험 데이터를 전처리 과정을 통해 불필요한 텍스트를 제거한 후, 비정형 데이터를 정형화 데이터로 변환한다. 라이브러리 gensim[16]에서 제공하는 Doc2Vec 모델을 사용하였고, 해당 Parameter에서 다음과 같이 설정하였다.

표 2. Doc2Vec Parameter
Table 2. Doc2Vec Parameter

Option	Value
DM	1
Vector_size	100
alpha	0.025
min_alpha	0.025
epoch	100

표 2의 옵션에서 DM을 1로 지정한 것은 해당 모델을 사용함을 의미한다, Vector_size는 벡터의 크기(dimension), alpha는 사용자가 지정하는 학습률(learning rate)를 의미하고, min_alpha는 최소 학습률을 의미한다. epoch는 모델의 훈련 횟수를 나타낸다. Doc2Vec 모델 Parameter에서 epoch를 100 설정한 만큼 학습된 DM모델이 만들어진다. 그 후 4개의 머신 러닝 기술을 각각 적용하여 분류 모델을 만든다.

3. 실험 결과

4개의 머신 러닝 기술을 사용하여 분류기를 각각 만든 후, 모델 성능 평가 지표로 비교 분석한다. 분류 모델의 성능 평가 지표는 크게 4가지로 평가된다. 정확도(Accuracy)와 정밀도(Precision), 재현율(Recall), 그리고 정밀도와 재현율의 조화 평균값인 F1 점수(F1-score)이다.

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+FN+FP+TN} \\ \text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN} \\ \text{F1}-\text{score} &= 2\,\frac{\text{Precision}\times Recall}{\text{Precision}+Recall} \end{aligned} \quad (1)$$

수식 1에서 TP(True Positive)는 위해도 관련 뉴스인 것을 제대로 검출된 것을, TN(True Negative) 그 외의 뉴스를 제대로 검출된 것이며, FP(False Positive)는 위해도 뉴스임에도 그 외의 뉴스로 잘 못 검출된 것으로, FN(False Negative)는 그 외의 뉴스임에도 위해도 뉴스로 잘 못 검출된 것임을 의미한다. 수식 1에서 표기된 식으로 4개의 분류 모델의 성능 평가 지표를 비교 분석하였다.

표 3 실험 결과로 정확도의 경우 로지스틱 회귀분석이 0.92로 가장 높은 정확도가 나왔고, 정밀도의 경우 결정 트리가 0.90으로 가장 높은 정밀도를 나타냈다. 재현율과 F1 점수의 경우 서포트 벡터 머신이 각각 0.82, 0.83으로 분류 모델의 성능 평가에서 F1 점수가 가장 높은 점수가 성능이 좋은 모델로써 평가하기에 4개의 머신 러닝 기술 중 서포트 벡터 머신이 가장 높은 성능을 나타냈다.

V. 결 론

본 연구는 머신 러닝 기술을 이용하여 과거의 수집된 문

표 3. 분류 모델 성능 비교 분석
Table 3. Classification Model Performance Comparison Analysis

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.92	0.85	0.80	0.82
Decision Tree	0.83	0.90	0.78	0.81
(Gaussian) Naive Bayesian	0.91	0.87	0.78	0.80
Support Vector Machine	0.88	0.85	0.82	0.83

서를 분석하고 이를 바탕으로 문서 분류 모델을 제안한다. 연구를 위해 도메인과 관련된 키워드를 기반으로 데이터를 수집하고, 특수문자와 같은 불용어를 제거한다. 그리고 형태소 분석기를 사용하여 수집한 문서의 각 단어에 품사를 태깅한다. 문서를 벡터로 변환하는 Doc2Vec 모델을 이용해 임베딩 모델을 만든 후 문서 간 유사도를 측정하고 머신러닝 기술을 이용하여 문서 분류기를 학습하였다.

연구에 사용할 생활 화학제품 관련 4800개의 뉴스 데이터를 Doc2Vec 모델로 학습시킨 후 4개의 머신 러닝 기술을 사용하여 분류기 성능을 비교 실험하였다. 실험 결과, 서포트 벡터 머신의 F1 점수가 0.83으로 가장 좋은 성능을 보였다.

향후에는 본 연구 결과를 기반으로 제품에 관련된 내용과 위해 정보의 포함 여부에 대해 라벨링하여 진행한 후 학습 데이터를 만든다. 그 후, 다양한 분야에 적용할 수 있는 다중 분류 모델 및 데이터의 계층적 표현을 위한 분류 모델을 연구한다.

참 고 문 헌 (References)

[1] Jun-Ho Roh, Han-joon Kim, Jae-Young Chang. "Improving Hypertext Classification Systems through WordNet-based Feature Abstraction." The Jounal of Society for e-Business Studies, 18.2 pp.95-110(6) 2013.May

[2] YunJeong Choi, SeungSoo Park. *"Interplay of Text Mining and Data Mining for Classifying Web Contents."* KOREAN JOURNAL OF COGNITIVE SCIENCE, 13.3 pp.33-46.(14) 2002.9

[3] Sunghae Jun *"A Big Data Preprocessing using Statistical Text Mining"* Journal of Korean Institute of Intelligent Systems Vol. 25, No. 5, pp. 470-476(7) 2015 October

[4] Eun-Soon You, Gun-Hee, Choi, Seung-Hoon Kim *"Study on Extraction of Keywords Using TF-IDF and Text Structure of Novels"* Korean Society of Computer Information Volume 20, Issue 2, pp.121-129(9) 2015 February

[5] J. Ramos, *"Using tf-idf to determine word relevance in document queries"*, In Proceedings of the First Instructional Conference on Machine Learning, 2003

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean *"Distributed Representations of Words and Phrases and their Compositionality"* NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 pp.3111-3119(9) Lake Tahoe, Nevada December 2013

[7] Garam Choi, Sung-Pil Choi *"A Study on the Deduction of Social Issues Applying Word Embedding: With an Empasis on News Articles related to the Disables"* Journal of the Korean Society for Information Management, 35(1) pp.231-250 (20) 2018.3

[8] Jung-Mi Kim, Ju-Hong Lee. *"Text Document Classification Based on Recurrent Neural Network Using Word2vec."* Journal of Korean Institute of Intelligent Systems, 27.6 pp. 560-565 (6) 2017.12

[9] Quoc Le ,Tomas Mikolov *"Distributed Representations of Sentences and Documents"* ICML'14 Proceedings of the 31st International Conference on International Conference on Machine Learning Volume 32 pp.1188-1196(9) Beijing, China June 2014

[10] Lucy Park, Sungzoon Cho, *"KoNLPy : Korean natural language processing in Python"* Proceeding soft he 26th Annual Conferenceon Human & Cognitive Language Technology, 2014 10

[11] Seong-Ho Choi, Eun-Sol Kim, Byoung-Tak Zhang *"An Intention Prediction Method for Dialogue using Paragraph Vector"* Korea Computer Congress 2016 pp.977-979(3) 2016.6

[12] KyuWan Kim, HyunJu Shin, SunJin Kim, KyoungDuek Moon, HyunAh Lee. *"Detecting Improper Paragraphs in a News Article Using Logistic Regression Classification and Inter-class Similarity."* Journal of Computing Science and Engineering pp.1873-1875.(3) 2017.12

[13] Dan-Ho Park, Won-Sik Choi, Hong-Jo Kim, Seok-Lyong Lee. *"Web Document Classification System Using the Text Analysis and Decision Tree Model."* Journal of Computing Science and Engineering, 38.2A 248-251.(4) 2011.11

[14] Do-Sik Min, Mu-Hee Song, Ki-Jun Son, Sang-Jo Lee. *"Spam - mail Filtering Using SVM Classifier."* Journal of Computing Science and Engineering 30.1B pp.552-554.(3) 2003.4

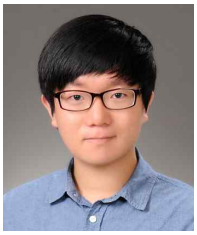
[15] Song-yi Han, Yong-Gyu Jung. *"Spam Filtering Using A Complement Naive Bayesian Classifier."* Journal of Computing Science and Engineering, 36.2C 325-328.(4) 2009.11

[16] scikit-learn, https://scikit-learn.org/stable/

저 자 소 개

정 지 수

- 2018년 : 숭실대학교 평생교육원 컴퓨터공학 학사
- 2018년 ~ 현재 : 세종대학교 소프트웨어융합학과 석사과정
- ORCID : https://orcid.org/0000-0003-3756-1074
- 주관심분야 : 텍스트 마이닝, 기계학습, 딥러닝



지 민 규

- 2018년 : 세종대학교 천문우주학과 학사
- 2018년 ~ 현재 : 세종대학교 소프트웨어융합학과 석사과정
- ORCID : https://orcid.org/0000-0002-3089-1452
- 주관심분야 : 텍스트 마이닝, 기계학습, 딥러닝

고 명 현

- 2016년 : 세종대학교 디지털콘텐츠학과 학사
- 2016년 ~ 현재 : 세종대학교 디지털콘텐츠학과 석사과정
- ORCID : https://orcid.org/0000-0002-6036-4717
- 주관심분야 : 텍스트 마이닝, 기계학습, 딥러닝

김 학 동

- 2016년 : 경성대학교 컴퓨터공학과 학사
- 2017년 ~ 현재 : 세종대학교 디지털콘텐츠학과 석,박사통합과정
- ORCID : https://orcid.org/0000-0003-3816-1224
- 주관심분야 : 머신러닝, 딥러닝, 자연어처리

임 헌 영

- 2017년 : 세종대학교 디지털콘텐츠학과 학사
- 2017년 ~ 현재 : 세종대학교 디지털콘텐츠학과 석사과정
- ORCID : https://orcid.org/0000-0002-8547-6248
- 주관심분야 : 컴퓨터 비전, 기계학습, 딥러닝

저 자 소 개

이 유 림

- 2018년 : 세종대학교 디지털콘텐츠학과 학사
- 2018년 ~ 현재 : 세종대학교 인공지능언어공학과 석사과정
- ORCID : https://orcid.org/0000-0001-8309-090X
- 주관심분야 : 텍스트 마이닝, 자연어 처리, 딥러닝

김 원 일

- 1981년 12월 ~ 1985년 7월 : ㈜대한항공 전산실 재무 시스템 개발원
- 1982년 : 한양대학교 공과대학 금속공학 학사
- 1987년 : 미국 일리노이주 서던일리노이대학교 컴퓨터 과학 학사
- 1990년 : 미국 일리노이주 서던일리노이대학교 컴퓨터 과학 석사
- 1994년 : 미국 인디아나주 인디애나 대학교 대학원 컴퓨터 과학 전공
- 2000년 : 미국 뉴욕주 시러큐스 대학교 대학원 컴퓨터 & 정보과학 공학 박사
- 2000년 1월 ~ 2001년 3월 : 미국 펜실바니아주 외인시 소재 Bhasha, INC Technical Staff (연구원)
- 2002년 3월 ~ 2003년 8월 : 아주대학교 정보통신전문대학원 BK 교수
- 2003년 9월 ~ 2017년 2월 : 세종대학교 전자정보공학대학 교수
- 2017년 3월 ~ 현재 : 세종대학교 소프트웨어융합대학 교수
- ORCID : https://orcid.org/0000-0002-1489-8427
- 주관심분야 : 인공지능, 지능형 시스템, 딥러닝 등