Word2vec과 CNN을 이용한 범죄 관련 뉴스기사 제목 분류 연구

장종욱*, 김동효*, 신현후*, 정준호*, 손윤식*
*동국대학교 컴퓨터공학과
e-mail : sonbug@dongguk.edu

A Study on the Title Classification of Crime-related News Articles Using Word2ec and CNN

Jong-Wook Jang*, Dong-Hyo Kim*, Hyunhu Shin*,
Junho Jeong*, Yunsik Son*
*Dept of Computer Science and Engineering, Dongguk Univ. Seoul, Korea

요 약

본 논문은 네이버 뉴스를 대상으로 웹 크롤러를 통해 수집된 범죄 관련 키워드 기반 뉴스기사 제목을 분석하기 위해 Word2vec과 CNN을 이용한 텍스트 형식의 뉴스기사 제목을 분류하는 방법론을 제안한다. 이를 통해 범죄 분석에 활용될 수 있는 의미 있는 뉴스기사 제목들을 선별할 수 있게 되어 범죄 관련 데이터를 효율적으로 분석할 수 있게 될 것으로 기대된다.

1. 서론

최근 범죄와 관련이 있는 데이터를 얻기 위해 웹 뉴스기사를 대상으로 웹 크롤링을 통해 빅데이터를 활용하여 범죄 관련 데이터를 분석하는 연구가 진행되고 있다. 크롤링된 범죄 관련 뉴스기사는 2018년 4월 11일부터 2019년 4월 10일까지 네이버 뉴스에서 범죄 키워드인 사기, 절도, 폭행, 손괴, 횡령, 상해, 추행, 협박, 성풍속, 도박, 마약, 강간, 공갈, 배임, 감금, 방화, 강도, 살인, 약취가 웹 뉴스기사 내용에 적어도 하나의 키워드가 포함된 뉴스기사를 말한다. 크롤러를 통해 수집된 뉴스기사는 차후에 범죄 예방을 위해 사용되는 범죄분석에 유용한 정보가 되기도 하지만, 그렇지 않은 뉴스기사도 있다. 예를 들어, '살인'이라는 키워드로 수집된 뉴스기사들 중에서 기사제목이 '나의 아저씨라는 드라마 명장면5' 에는 "동훈(이선균)은 살인자가 될 수 밖에 없었던 지안(이지은)의 불우한 과거를 알고도 등 돌리지 않았다"라는 기사내용이 포함되어 있어서 '살인'이라는 키워드로 범죄 관련 기사로 크롤링 된 것이다. 하지만 이러한 뉴스기사는 범죄 예방을 위한 범죄 분석에 필요한 정보가 아니기 때문에 범죄 관련 기사제목을 분류해주는 작업이 필요하다. 이렇게 드라마, 영화, 연예인 등과 같은 기사 중에서 위에서 제시한 범죄 관련 키워드 중심으로 기사들이 크롤링되어 기사가 수집될 수 있으며, 범죄 분석에 유용한 범죄와 관련된 뉴스기사 제목을 분류하여 범죄 관련 의미 있는 데이터를 구성할 필요가 있다. 이를 위해 본 논문에서는 범죄 분석에 불필요한 데이터들을 분류하고, 범죄 분석에 유용한 데이터만을 선별하여 범죄 키워드 기반 빅데이터를 구성할 수 있는 방안을 제안한다. 2장에서는 기존의 연구들을 소개하고, 3장에서는 분류모델을 제안하며, 4장에서는 결론 및 향후 연구에 대해 다룬다.

"이 성과는 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행되는 연구임 (No.2018R1A5A7023490).

2. 관련연구

2.1 단어 임베딩(Word Embedding)

단어 임베딩(Word Embedding)은 각 단어가 분산된 의미를 지니도록 고정된 차원의 벡터공간에 대응 시키는 방법으로, 기계번역, 개체명 인식 등 많은 자연어처리 분야에서 활용되고 있으며[1], 단어 임베딩의 대표적인 방법으로는 word2vec모델이 있다[4]. word2vec은 특정 embedding 공간상에서 같은 맥락(context)을 갖는 단어들이 가까운 거리를 가진다는 전제에서 출발한다. 이러한 word embedding 방식의 word2vec 표현법은 주어진 문장에 대한 문법적 해석이 가능하며, 단어의 거리를 통해 의미론적 추론도 가능하다[2].

2.2 형태소 분석(Morphological analysis)

한국어의 경우 교착어에 속하는 특성으로 인해 문장에서 어근이 독립적으로 존재하지 않는다. 또한 60여 가지에 이르는 조사 및 어미로 인해 다른 어절이 동일한 형태소를

포함하는 경우가 많다. 예를 들어. “집에서” 와 “집으로”는 동일한 “집”이라는 형태소를 지니고 있지만 각각 다른 조사가 뒤에 붙음으로서 다른 형태의 어절로 나타낸다. 따라서 영어 같이 띄어쓰기 단위를 한국어 단어 임베딩 학습의 입력으로 사용한다면, 하나의 어근에 대한 너무 다양한 형태가 존재하여 어근 자체가 지니는 임베딩을 학습하기 어렵다. 또한 많은 형태에 따라 비례하여 늘어나는 단어 수는 단어 당 학습에 사용되는 컨텍스트(context words)의 부족으로 이어져 학습 성능을 떨어뜨리는 요인이 된다.

형태소 분석기는 주어진 문장을 형태소 단위로 나누어 어근과 조사, 어미 등을 분리해주고 어근 역시 명사 및 동사, 형용사 등으로 구분시켜준다. 대표적인 형태소 분석기로는 트위터 형태소 분석기와 꼬꼬마 형태소 분석기가 있으며, 이를 이용해 말뭉치에 등장하는 모든 문장에서 독립적으로 의미를 지니지 못하는 형식형태소를 제거하고, 체언과 용언 등 실질적인 의미를 지니는 단위를 단어 임베딩의 입력단위로 사용하여 단어 임베딩 학습이 가능하다[1].

2.3 컨볼루션 신경망(Convolutional Neural Network)

컨볼루션 신경망은 사람의 신경망에서 고안한 모델로 다양한 패턴인식 문제에 사용되고 있다. 두 가지 연산 층(convolutional, subsampling 혹은 max-pooling 층)을 번갈아 수행하며, 최종적으로는 fully connected layer를 통해 분류를 수행하는 계층 모델이다. 텍스트에서 컨볼루션 신경망의 첫 번째 층은 문장 속 단어들을 테이블 lookup을 이용하여 단어 벡터로 만든다. 각 단어를 픽셀로 생각하고 각 문서를 문서 당 단어의 개수만큼의 채널을 가진 (|문서|)×1 벡터로 표현한 뒤 그 나머지는 이미지의 경우와 동일하게 단어 벡터를 feature 벡터로 사용한다[3].

3. 범죄 관련 뉴스기사 제목 분류모델

본 논문에서는 웹크롤러로부터 수집된 텍스트 형식의 네이버 뉴스기사의 제목(title)을 분류하는 모델을 제안한다. 예를 들어 크롤러로부터 수집된 뉴스기사 제목 또는 내용에 ‘횡령’ 키워드가 포함된 뉴스기사들 중에서 드라마, 영화, 연예 등 범죄와 관련이 없는 기사 제목들을 분류하여 범죄 분석을 위한 의미 있는 뉴스기사를 제공해주는 모델을 제안한다. 아래 그림2는 본 논문에서 제안하는 모델의 구조로서 모델의 입력은 텍스트 형식의 뉴스기사 제목과 해당 제목이 범죄 관련 뉴스기사인지 아닌지 분류해주는 값인 1과 0이다. 이때 1은 범죄 관련 뉴스기사 제목임을 나타내고, 0은 드라마, 영화 등과 같은 범죄와는 무관한 기사제목임을 나타낸다. 텍스트와 분류값은 형태소 분석을 통해 명사로 추출하고 word2vec은 딥러닝 분류기의 학습을 위해 텍스트를 벡터값으로 변환 후 딥러닝 분류기를 통해 학습데이터와 평가데이터를 학습시킨다. 그리고 모델의 출력으로 확률값을 표현하여 0.5이상이면 범죄 관련 기사이고, 0.5 미만이면 범죄 관련 기사가 아닌 드라마, 영화, 연예 등과 관련된 기사로 분류할 수 있게 된다.

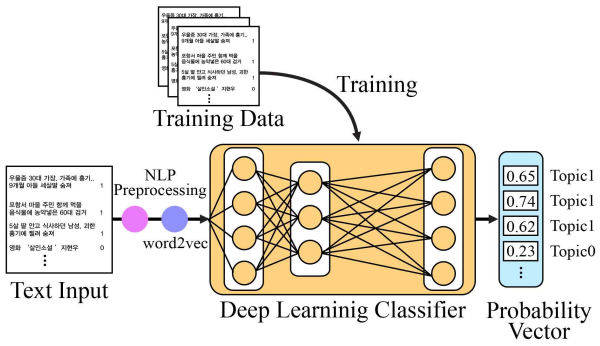

그림 2. 뉴스기사 제목 분류모델 구조

3.1 실험 데이터

본 실험에서 사용된 데이터는 앞 절에서 설명했듯이 네이버 뉴스기사를 대상으로 실험을 했으며, 크롤링되어 수집된 뉴스기사는 MongoDB에 저장되어 있고, 기사가 수집된 기간은 2018년 4월 11일부터 2019년 4월 10일까지이다. 또한 네이버 뉴스에서 범죄 키워드인 사기, 절도, 폭행, 손괴, 횡령, 상해, 추행, 협박, 성풍속, 도박, 마약, 강간, 공갈, 배임, 감금, 방화, 강도, 살인, 약취가 웹 뉴스기사 제목(title) 또는 내용(content)에 위에서 제시한 범죄 키워드가 적어도 하나 이상 포함된 뉴스기사를 말한다. 범죄 키워드별 수집된 기사건수는 총 약 70만 5천건이고, 그 중 살인은 약 10만 3천건, 폭행은 약 15만건, 횡령은 약 5만 3천건의 뉴스기사가 수집되었다.

3.2 데이터 전처리

띄어쓰기, 의미를 지니는 체언과 용언을 기준으로 문서를 토큰화하여 특수 문자가 제거된 단어로 문서를 재구성하였다. 그리고 토큰화된 단어를 벡터로 표현하기 위하여 word2vec을 활용하였다. word2vec을 활용하면 의미가 유사한 단어나 문법적으로 비슷한 구조를 이루는 단어는 embedding 공간 상 가까운 벡터 공간에 놓이게 된다. 즉, 문서를 표현하는 단어의 벡터 값들이 범주별로 군집화 되어 다른 범주에 속한 문서들과 벡터 공간상의 위치에 있어 구분될 수 있도록 단어의 벡터 표현이 생성된 것이다[2]. 컨볼루션 신경망의 학습에 앞서 이러한 데이터 전처리 방식은 자질들의 변별적인 특징을 잘 내포하고 있는 벡터 값으로 단어를 수치화하여 학습 성능을 높일 수 있다.

3.3 제안 모델 알고리즘

한국어 형태소 분석기를 사용하기 위해 파이썬 패키지인 KoNLPy를 사용하였고[5], 텐서플로(TensorFlow)를 실험에서 수행하기 위해 파이썬으로 작성된 오픈 소스 신경망 라이브러리인 케라스(Keras)를 사용하려고 한다. 향후 전체 실험 데이터 중 기사의 내용에 ‘살인’이라는 키워드가 포함된 기사 제목으로 실험분석을 하려고 한다. 본 논문에서 제안하는 모델 알고리즘은 다음과 같다.

표1. word2vec 모델 알고리즘

조사, 어미, 문장부호를 제외한 단어들을 입력으로 하여 모델을 만드는 과정: 1. 100차원으로 벡터 차원 설정 2. 앞 뒤 10개의 주변 단어를 문맥으로 보고 예측 모델에 반영 3. 소프트맥스 사용, 단어 출현빈도 2번 미만은 제외시킴 4. word2vec의 cbow와 skip-gram알고리즘 중 skip-gram 사용

표2. 텍스트를 벡터로 변환 알고리즘

학습용, 검증용 데이터셋 대상으로 명사만 추출하고 word2vec모델을 불러와서 텍스트를 벡터로 변환하는 과정: 5. 성능을 높이기 위한 작업으로 학습용 데이터셋과 검증용 데이터셋 단어들을 대상으로 명사만 추출 6. 텍스트의 길이가 짧은 경우 길이를 맞추기 위해 패딩(padding)처리 7. 명사형 단어들을 벡터모델에 적용 8.. word2vec모델에 단어가 없는 경우, 새로운 단어에 대한 예외처리 적용

표3. 컨볼루션 신경망 모델 알고리즘

컨볼루션 신경망 모델을 만드는 과정: 9. 입력층의 데이터 (입력 가로 사이즈, 벡터차원 등 설정) 10. 첫번째 Conv 레이어 (3x3 필터 갯수 1개, 필터가 움직일 때 x축으로 1, y축으로 1) 11. 두번째 Conv 레이어(위에서 만든 값이 아래 값에 적용되도록 함) 12. fully connected network (Dropout 0.5) 13. 출력계층 생성 (활성화 함수는 시그모이드 함수를 사용 (0.0~1.0사이의 값으로 출력됨) 14. CNN모델 생성 15. 모델 컴파일, 손실함수, 최적화기 설정 (Adam 최적화기로 정확도 측정 (학습률은 0.001로 설정)

표4. 학습 및 성능평가 알고리즘

학습을 시키고 학습 데이터셋과 평가 데이터셋 성능평가 과정: 16. 학습 모델의 타당성 검증을 위해, 전체 데이터 중 25%는 검증용 데이터로 사용함 17. 한 번에 벡터를 256개 처리(batch size:256), 전체 데이터에 대한 학습 횟수 5번(epochs:5)으로 설정 18. 학습을 할 때 성능이 떨어지는 경우, 두 번이상 정확도가 떨어지면 멈추는 학습 조기종료 콜백함수 추가 19. 학습 데이터셋, 검증용 평가 데이터셋 손실률과 정확도 측정

4. 결론 및 향후연구

본 논문에서는 크롤링을 통해 범죄 키워드 기반으로 수집된 뉴스기사의 제목(title)을 분류 및 정제 해주는 모델을 제안하였다. 이를 통해 기사 제목을 기준으로 범죄 관련 뉴스기사를 수집하는 과정에서 범죄 예방 및 범죄 분석에 유용한 정보를 얻는데 도움이 될 것으로 기대된다. 향후 뉴스기사 제목 분류모델 정확도를 향상시키는 연구뿐만 아니라 기사 본문 내용도 분석하여 범죄 관련 뉴스기사 본문에 대한 분류 정확도를 향상시킬 수 있는 방법론을 연구 할 예정이다.

참고문헌

[1] 최상혁, 설진석, 이상구 “한국어에 적합한 단어 임베딩 모델 및 파라미터 튜닝에 관한 연구,” 한글 및 한국어 정보처리 학술대회, pp. 252, 2016.

[2] 김정미, 이주홍 “Word2vec을 활용한 RNN기반의 문서 분류에 관한 연구,” 한국지능시스템학회, 제27권, 6호, pp. 560-565, 2017.

[3] 조휘열, 김진화, 윤상웅, 김경민, 장병탁 “컨볼루션 신경망 기반 대용량 텍스트 데이터 분류 기술,” 한국정보과학회, pp. 792-794, 2015.

[4] KIM. Yoon, “Convolutional neural networks for sentence classification,” *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746-1751, 2014.

[5] 박은정, 조성준, “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 제 26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.