텍스트 마이닝을 이용한 기사 내 부적합 문단 검출 시스템

김규완[0], 신현주, 김선진, 이현아
금오공과대학교, 컴퓨터소프트웨어공학과
rla9826@naver.com, dotcomehe@naver.com, junnis0123@naver.com, halee@kumoh.ac.kr

Detecting Improper Sentences in a News Article Using Text Mining

[Kyu-Wan Kim[O], Hyun-Ju Sin, Seon-Jin Kim. Hyun Ah Lee]
Dept. of Computer Software Engineering, Kumoh National Institute of Technology

요 약

SNS와 스마트기기의 발전으로 온라인을 통한 뉴스 배포가 용이해지면서 악의적으로 조작된 뉴스가 급속도로 생성되어 확산되고 있다. 뉴스 조작은 다양한 형태로 이루어지는데, 이 중에서 정상적인 기사 내에 광고나 낚시성 내용을 포함시켜 독자가 의도하지 않은 정보에 노출되게 하는 형태는 독자가 해당 내용을 진짜 뉴스로 받아들이기 쉽다. 본 논문에서는 뉴스 기사 내에 포함된 문단 중에서 부적합한 문단이 포함되었는지를 판정하기 위한 방법을 제안한다. 제안하는 방식에서는 자연어 처리에 유용한 Convolutional Neural Network(CNN)모델 중 Word2Vec과 tf-idf 알고리즘, 로지스틱 회귀를 함께 이용하여 뉴스 부적합 문단을 검출한다. 본 시스템에서는 로지스틱 회귀를 이용하여 문단의 카테고리를 분류하여 본문의 카테고리 분포도를 계산하고 Word2Vec을 이용하여 문단간의 유사도를 계산한 결과에 가중치를 부여하여 부적합 문단을 검출한다.

주제어: deep learning, text embedding, Word2Vec, Doc2Vec, logistic regression, 부적합 문단 검출

1. 서론

인터넷의 발전으로 인해 뉴스 콘텐츠는 TV보도와 종이 신문을 대신하여 인터넷 뉴스가 주류를 이루는 형태로 변화했다. 이로 인하여 뉴스 독자는 개인의 필요와 관심에 맞는 뉴스 기사를 선택하여 읽을 수 있게 되었다. 하지만 사람들이 흥미 위주로 뉴스 기사를 열람하는 성향이 강해지면서 이를 악용하여 가짜 뉴스를 배포하는 개인 및 기업이 등장했다. 특히 SNS와 개인 미디어를 통하여 확산되는 가짜 뉴스는 사람들을 속이기 위한 목적으로 만들어진 만큼 그 파급력이 크다[1]. 이러한 가짜 뉴스가 점차 늘어나면서 사람들의 혼란이 초래되고 매체에 대한 불신이 높아지고 있다. 하지만 수많은 매체가 쏟아내는 방대한 양의 뉴스 데이터를 사람의 손으로 분류해내는 데에는 비용과 시간의 한계가 있다. 때문에 가짜 뉴스를 자동으로 분류해내기 위한 알고리즘에 대한 연구는 필수 불가결이다. 하지만 가짜 뉴스가 무엇인지 명확하게 정의하기 어렵고, 사람조차 구분해내기 힘든 사실에 대한 진실과 거짓을 컴퓨터가 분류하는 것은 쉽지 않은 문제이다.

이러한 가짜 뉴스의 한 유형에는 화제성 높은 주제를 채택하고 본문 내부에 다른 분야의 문단을 섞어 넣거나 광고성 문구를 포함하는 것이다. 본 논문에서는 이러한 가짜 뉴스 유형을 본문에 해당되는 카테고리와 다른 카테고리의 문단이 포함된 뉴스로 보고 이러한 문단을 검출하기 위한 방법을 제안한다. 뉴스 문단의 카테고리의 자동 분류에서는 자연어 처리에 대해 이미 성능이 검증된 Word2Vec 방식을 적용하고, tf-idf 가중치 알고리즘에 로지스틱 회귀를 적용한 뉴스 부적합 문단 검출 방법을 제안한다.

2. 관련 연구

최근 들어 Convolutional Neural Network(CNN)가 자연어처리에 적용되기 시작하면서 놀라운 결과를 얻고 있다[2]. CNN모델 중 하나인 Word2Vec은 입력한 말뭉치의 문단에 있는 단어와 인접 단어의 관계를 이용하여 단어의 의미를 학습한다. Word2Vec의 학습 방법은 CBOW, Skip-gram의 두 종류가 있다. CBOW(Continuous Bag Of Words)방식은 주변 단어가 만드는 맥락을 이용해 타겟 단어를 예측하고, Skip-gram은 한 단어를 기준으로 주변에 올 수 있는 단어를 예측한다.

Word2Vec[3]의 학습 과정은 큰 틀에서 일반적인 인공신경망의 학습과 비슷하다. 한 단어에 이미 할당된 벡터, 즉 단어 임베딩(word embedding)이 있다고 가정하고 이 값을 이용해 주변 문맥을 얼마나 정확하게 예측하는

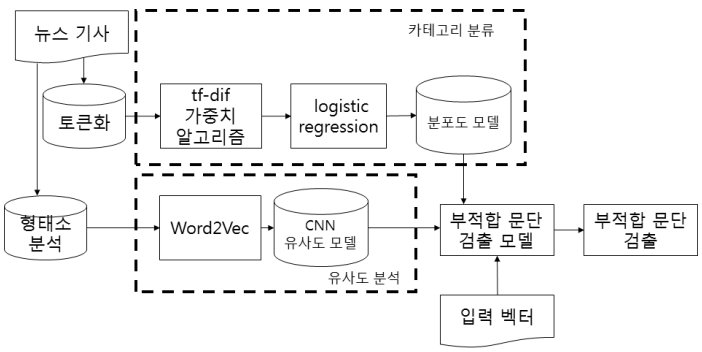
지 계산한다. 학습 과정에서 한 단어를 기준으로 단어 주변의 문맥을 참고하여 현재 임베딩 된 벡터가 얼마나 정확한지, 오차의 값은 어느 정도인지를 알아낸다. 만일 어떤 두 단어가 비슷한 문맥에서 꾸준하게 사용될 경우 두 단어는 비슷한 벡터 값은 갖는다. 학습이 잘 완료되었다면 고차원 공간에서 비슷한 단어가 근처에 위치하게 되는데 이 관계를 이용하면 유사성을 알아낼 수 있다.

Doc2Vec[4]은 Word2Vec을 문단, 단락 또는 전체 문서와 같이 더 큰 텍스트 블록에 대한 연속 표현을 학습하도록 수정한 모델이다. Doc2Vec은 단어와 레이블에 대한 표현을 동시에 학습한다. 본 논문에서는 카테고리 분류와 유사도 모델을 혼합하여 부적합 문단을 추출하고자 하는데, 카테고리 분류에 대해서는 Doc2Vec과 Word2Vec을 혼합하여 분류율을 비교한 김도우[4]의 연구가 있다. 또한, Word2Vec을 이용하여 문서간 유사도를 비교한 사례가 있다[5].

3. 제안 시스템

본 논문에서는 Word2Vec을 통한 word Embedding을 통한 유사도 모델과, 공백으로 토큰화 한 후 tf-dif 가중치 알고리즘과 Logistic regression를 적용한 분포도 모델을 통하여 부적합 문단을 검출하는 모델을 제안한다.

제안 모델의 구조는 아래 [그림 1]과 같으며, 시스템은 크게 카테고리 분류 부분과 유사도 분석 부분으로 나뉜다.



[그림 1] 시스템 구조도

문서에 비해 단어의 수가 상대적으로 부족한 문단 단위의 분류를 무리 없게 수행하기 위해서는 충분한 데이터 셋의 구축이 필요하다. 본 실험에서는 조선일보에서 2017년 8월 ~ 2007년 3월 일자까지 수집한 뉴스 데이터를 사용했다. 학습 데이터 셋은 각 카테고리 별 5만 건의 뉴스 기사로 이루어진 30만 건의 뉴스 데이터로 구축하였다.

분포도 모델의 학습에 사용 될 레이블은 뉴스 카테고리를 [0]스포츠, [1]정치, [2]경제, [3]사회, [4]연예/방송, [5]오피니언/칼럼/사설로 나누어 0~5의 식별자를 부여했다. 학습된 분포도 모델에 입력 벡터를 입력하면 각 문단에 대한 레이블을 반환받을 수 있다. 여기서 입력 벡터는 기사 제목, 본문으로 구성된다. 제안 모델에서는 카테고리 분류를 위하여 각 문단에서 반환된 레이블과 문단의 수를 이용한 분포도를 계산한다. 예를 들어, 분포도 모델에서 입력 벡터를 분석한 결과가 아래 [그림 2]와 같다고 가정하자. 부적합 문단을 검출하기 위해서는 본문의 레이블 분포 파악이 필요하다. 우선 시스템에서는 각 레이블에 포함되는 문단의 개수를 전체 문단의 개수로 나누어 레이블별 분포율을 구한다. [그림 2]의 예제에서 총 문단 수는 4개이고, 그 중 레이블이 0인 문단이 3개, 레이블이 4인 문단이 1개이다. 그러므로 [0]스포츠에 대한 분포율은 0.75, [4]연예/방송에 분포율은 0.25가 되고, 문단 A와 B, C는 0.75, 문단 D는 0.25의 분포율을 가진다.

손흥민(토트넘)이 유럽축구연맹(UEFA)챔피언스리그에서 시즌 첫 골을 폭발시켰다. [문단A:레이블0]
손흥민은 14일(한국시간) 오전 영국 런던 웸블리 스타디움에서 열린 2017-2018시 UEFA 챔피언스리그 조별리그 1차전 도르트문트(독일)와 홈 경기에서 득점포를 가동했다. [문단B:레이블 0]
이날 선발 출격한 손흥민은 0-0이던 전반 4분 하프라인 아래에서 해리 케인의 패스를 받은 뒤 도르트문트의 왼쪽 진영을 뚫었다. [문단C:레이블 0]
'군함도'는 영화 '베테랑'을 만든 류승완 감독 작품이다. 송중기, 황정민, 소지섭, 이정현 씨 등 인지도 높은 배우들이 출연해 개봉 전부터 주목받았다. [문단D:레이블 4]

[그림 2] 가짜 뉴스의 예시

분포도 모델은 토큰화된 뉴스 데이터를 tf-dif 가중치 알고리즘을 적용하여 임베딩된 벡터와 레이블을 Scikit-learn library[6]에서 제공하는 로지스틱회귀분석(logistic regression) 모델을 통해 학습을 하였다. 김도우의 실험을 참고하면 Doc2Vec 모델만을 이용했을 때, 학습되는 data와 vector의 차원을 조절하여도 카테고리 분류율이 고정되는 문제점이 있다. 이에 본 연구에서는 선형 분류 문제를 해결하기 위해 단순하면서도 보다 강력한 분류 알고리즘인 로지스틱회귀분석을 함께 사용하였다. 우선, 토큰화 과정에서 Doc2Vec은 twitter 형태소 분석기로 추출한 명사만을 학습하였다. Doc2Vec을 이용하여 임베딩 된 Vector와 레이블을 로지스틱회귀분석 모델로 학습시켜 얻어낸 분류율은 83.6%로 나타났다. 그러나 뉴스 기사의 경우 명사 외의 형태소도 카테고리 분류에 영향을 미칠 수 있다. 그에 따라 공백으로 토큰화하여 얻은 모든 형태소를 tf-dif 가중치[7] 알고리즘으로 추출한 Vector와 레이블을 로지스틱회귀분석 모델로 학습시켜 87.0%의 카테고리 분류율을 얻을 수 있었다.

유사도 모델은 형태소 분석된 뉴스 기사 본문 각 문단

에 존재하는 명사 벡터 사이의 유사성을 기준으로 문단의 유사도를 계산한다. 입력받은 각 n개의 문단에 대하여 대상 문단을 제외한 다른 문단과의 유사도를 구하고, 얻어진 유사도들의 평균값을 각 문단의 유사도로 부여한다. 예를 들어 [그림 1]에서 문단 1과 문단2의 유사도를 sim(1,2), 문단1과 문단3의 유사도를 sim(1,3), 문단1과 문단4의 유사도를 sim(1,4)라 하면, 문단 1의 평균 유사도는 sim(1,2)+sim(1,3)+sim(1,4)/3이 된다.

제안하는 시스템에서는 분포도 모델과 유사도 모델을 통해 나온 문단별 분포도와 유사도를 결합하여 적합도를 계산한다. 적합도 = 분포도×유사도이며, 본문에서 적합도가 가장 낮은 문단을 부적합 문단으로 판단하기로 한다.

예를 들어, [그림 2]의 예제를 이용하여 유사도 모델을 적용한 결과 값은 아래 [그림 3]과 같다.

	문단A	문단B	문단C	문단D
문단A	1	0.12	0.11	0.06
문단B	0.12	1	0.14	0.07
문단C	0.11	0.14	1	0.06
문단D	0.06	0.07	0.06	1

[그림 3] [그림 2]예제에 대한 유사도 실험 결과

분포도 모델에서는 각 문단에 대하여 0, 0, 0, 3의 레이블을 반환받았으며, 각 모델에서 얻은 분포도와 유사도로 도출한 적합도는 문단A의 경우 0.72, 문단B의 경우 0.825, 문단C의 경우 0.77, 문단D의경우 0.15이다.

위 실험 결과를 통해 제안 모델에서는 부적합한 문단의 적합도가 상대적으로 낮게 나오는 것을 알 수 있다. 그러나 적합도가 낮다고 해서 모두 부적합 문단으로 간주하기는 어렵다. 그러므로 본 논문에서는 문단별 부적합도를 측정하여 결과값을 출력한다.

4. 실험 및 결과

4.1 실험 데이터 구성

테스트 데이터 셋은 각 카테고리에 대해 조선일보에서 2017년 9월 뉴스를 수집한 뒤 부적합 문단이 포함되지 않은 1만 건의 데이터와 임의로 타 카테고리 기사의 한 문단을 포함시킨 1만 건의 데이터로 구축하였다. 부적합 문단이 포함되지 않은 뉴스는 43218개의 문단, 부적합 문단이 포함 된 문단은 43214개로 이루어져 있다.

4.2 학습 결과

유사도 모델의 경우 각 문단에 포함된 모든 명사간의 유사도를 조사한 평균값을 문단 간의 유사도로 설정하였다. 86432개의 테스트 셋에 대하여 평균 72%의 정확도를 나타냈다. tf-dif 가중치 알고리즘을 적용하여 임베딩된 벡터에 대해 scikit-learn 지도학습을 활용한 분포도 모델의 분류율은 문단에 대한 레이블 반환 값에 대한 단순 비교로 측정하였으며, 86432개의 테스트 셋에 대하여 평균 87.0%의 정확도를 나타냈다. 반환 레이블에 대해서는 모든 레이블과 비교하여 가장 높은 확률을 기록한 레이블을 채택하였다.

4.3 평가 결과 분석

본 시스템의 정확도를 평가하기 위해 검출 분야에서 사용되고 있는 평가방식으로 정확성(Accuracy)을 이용하여 제안된 알고리즘의 성능을 평가한다. 본 시스템의 실험 결과는 다음 [표 1]와 같다.

[표 1] 테스트 셋에 대한 실험 결과

		label	
		True	False
Prediction	True	TP : 38896건	FP : 4422건
	False	FN : 4322 건	TN :38792건

본 시스템의 Accuracy(정확성)는 $\frac{tp+tn}{tp+tn+fp+fn}$ 수식을 통하여 89.88%의 결과를 볼 수 있다.

5. 결론

본 논문에서는 가짜 뉴스에서 부적합한 문단 검출을 위해 CNN모델의 Word2Vec과 공백을 이용한 토큰화, tf-dif 알고리즘을 적용하여 로지스틱 회귀를 이용한 새로운 방법을 시도했다. Doc2Vec을 통해 부적합 문단을 검출하는 방식만으로는 높은 정확도를 기대할 수 없었던 반면, 토큰화와 tf-dif 알고리즘과 로지스틱 회귀를 사용함으로써 기사의 불특정한 부분에서 나타나게 되는 부적합 문단 검출에 높은 성능을 기대 할 수 있게 되었다. 로지스틱 회귀를 이용한 문단별 카테고리 분류와 Word2Vec을 이용한 문단 간 유사도에 가중치를 부여하여 얻는 적합도 점수를 통한 부적합 문단 검출은 각 모델을 단독 사용했을 때보다 시스템 성능이 좋게 나타났다. 그리고 학습을 계속 진행할수록 결과가 좋게 나오는 것을 볼 수 있었다. 또한, 제안 모델은 본 논문에서 주제로 삼은 가짜 뉴스 부적합 문단 검출뿐만 아니라 자기소개서나 논설문 등 부적합한 문단이 포함될 수 있는 글 또는 시스템 전반에 적용 가능할 것으로 보인다.

향후 연구로는 Word2Vec의 성능 향상을 위한 모델링 개발과 RNN 연구를 함께 진행할 것이다.

참고문헌

[1] 권만우, 전용우, 임하진, "가짜뉴스(Fake News) 현황분석을 통해 본 디지털매체 시대의 쟁점과 뉴스콘텐츠 제작 가이드라인", 멀티미디어학회 논문지, 제18권, 제11호, 1419-1426, 2015

[2] Ye Zhang, Byron C. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification", arXiv:1510.03820, 2015.

[3] Yoon Kim, "Convolutional Neural Network for Sentence Classification", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP), 2014.

[4] 김도우, "Doc2Vec을 활용한 CNN 기반 한국어 신문기사 분류에 관한 연구", 제28회 한글 및 한국어 정보처리 학술대회 논문집(2016년), 2016

[5] 황명권, 공현장, 황광수, 김판구. "문서의 계층화를 이용한 문서비교 방법." 한국정보과학회 학술발표논문집, 33.2B (2006.10): 143-147.

[6] http://scikit-learn.org/stable

[7] https://ko.wikipedia.org/wiki/TF-IDF