딥러닝 기법을 이용한 낚시성 기사 제목 분류에 대한 연구*

최용석[0], 최한나, 신지혜, 정창민,
안정연, 유채영, 임채은, 이공주
충남대학교
yongseok.choi.92@gmail.com, hannach1105@naver.com, shyen.jh@gmail.com, como411s@hanmail.net,
djqjdlskf333@naver.com, ycy4595@naver.com, collcr@naver.com, kjoolee@cnu.ac.kr

A study on classification of hooking headlines using deep learning techniques*

Yong-Seok Choi[0], Han-Na Choi, Ji-Hye Shin, Chang-Min Jeong,
Jung-Yeon An, Chae-Young Yoo, Chae-Eun Im, Kong-Joo Lee
Chungnam National University

요 약

본 논문은 낚시성 기사 제목과 비낚시성 기사 제목을 판별하기 위한 시스템을 제시한다. 서포트 벡터 머신(SVM)을 이용하여 기사 제목을 분류하며, 분류하는 기준은 딥러닝 기법중의 하나인 워드임베딩(Word Embedding), 군집화 알고리즘 중 하나인 K 평균 알고리즘(K-means)을 이용한다. 자질로서 기사 제목의 단어를 사용하였으며, 정확도가 83.78%이다. 결론적으로 낚시성 기사 제목에는 낚시를 유도하는 특별한 단어들이 존재함을 알 수 있다.

주제어: SVM, 낚시성 기사 제목, 기사분류, 단어 임베딩

1. 서 론

낚시성 기사 제목이란 '뉴스 내용에 상관없이 독자로 하여금 호기심을 자극하여 오로지 해당 뉴스에 대한 클릭을 유도할 목적으로 편집된 제목'으로 정의할 수 있다[1]. '요가 강사 ○○○'의 여러 TV 프로그램에서의 활동에 대해 소개하는 신문기사의 제목이 '요가 강사 ○○○, 타이트한 의상 속 아찔한 애플힙 눈길'로 지어지는 경우가 그 예이다. 이러한 기사는 제목에서 사용한 자극적인 단어로 사람들이 기사를 클릭하도록 유도한다. 기사를 클릭한 독자는 호기심을 일으킨 제목과 관련한 내용을 기대하지만 제목과 내용적인 연관성이 낮은 기사내용을 확인하게 된다. 최근 이러한 과정이 반복되면서 인터넷 신문기사의 신뢰성을 떨어뜨리는 문제가 발생하고 있다. 본 논문에서는 이러한 문제로부터 착안하여, 독자가 내용을 읽지 않고도 낚시성 기사 제목 분류를 도와주는 시스템에 대한 연구를 수행하였다.

본 연구와 관련된 선행 연구로는 기사 제목이 어느 정도 낚시성편집이 되고 있는지의 실태를 논한 사례가 있다[2]. 본 연구는 기존의 실태를 확인하는 데에 그치는 것이 아니라, 낚시성 기사 제목에 대해 분류하는 시스템을 제시한다.

2. 본 론

2.1 낚시성 기사 제목 분류

낚시성 기사 제목 분류를 위한 학습과 평가를 위해 인터넷 포털로부터 신문기사 제목과 함께 신문기사 본문도 함께 수집하였다. 수집한 기사의 낚시성, 비낚시성에 대한 분류는 6명의 인원이 참여하여 1차로 분류하고, 2명의 평가자가 교환하며 분류된 기사를 재검토하는 방법으로 낚시성 기사 제목을 분류하였다.

분류한 낚시성 기사 제목의 예는 [표 1]과 같다.

[표 1] 낚시성 기사 제목 예

제목
A, B 또? 속옷만 입고...
A 걸그룹 퇴출 통보, 결혼설 때문?

2.2 기본 자질 추출

기사의 전체 단어에 대해 한 기사에 나온 단어의 유무를 바이너리 값으로 자질 벡터를 구성하거나 빈도수를 자질로 사용하게 되면 매우 희박한 자질벡터가 만들어진다. 그러므로 본 논문에서는 딥러닝 기법 중에 하나인 워드 임베딩(Word Embedding)[3]을 이용해서 기사 제목에 나온 단어의 임베딩 벡터의 각 성분의 평균을 계산하여 자질벡터를 구성한다.

워드 임베딩

임베딩 모델 생성에는 인터넷에서 수집한 신문 기사와 불건전한 낚시성 기사를 신고하는 옴부즈맨카페[4]의 기사 제목, 나무위키**의 텍스트를 합친 말뭉치를 사용하였다.

신문기사와 일반 문서를 함께 사용함으로써, 신문 기사 도메인과 함께 실제 우리가 사용하는 일상 언어들과 다

* 이 논문은 2015년도 교육부의 재원으로 한국과학창의재단의 지원을 받아 수행된 연구임.

** https://namu.wiki/

양한 주제의 상세한 내용으로 다른 특성을 임베딩할 수 있도록 하였다.

워드 임베딩 모델 생성을 위해 Word2Vec[3] 알고리즘을 사용하였다. [표 2]는 임베딩 된 데이터에서 '바보'라는 단어를 검색하였을 때 가장 유사한 단어를 뽑은 결과이다.

[표 2] 워드 임베딩 결과

단어	유사도
멍청이	0.692436
멍청	0.642021
어린애	0.606459
변태	0.588238
겁쟁이	0.581658
구제불능	0.558232
바보짓	0.553802
게으름뱅이	0.546978
꼬맹이	0.545882
츤데레	0.544073
병신	0.542776

2.3 자질 확장

수집된 기사로부터 추출한 단어는 기사의 제목이 짧기 때문에 단어의 개수가 한정적이다. 그래서 내용에 있는 제목과 유사한 단어를 실험 제목의 확장의 개념에서 사용하기 위해 K 평균 알고리즘을 활용하였다.

군집화

K 평균 군집화 알고리즘은 군집화 기법 중에 하나이며 임의로 설정한 k개의 군집들을 만들어 유사한 단어나 문장을 하나로 표현해주는 방법이다[5]. 기사제목의 단어 수가 내용에 비해 상대적으로 부족하여 더 나은 실험결과가 도출되기 어렵다고 판단하였다. 그렇기 때문에 제목과 내용의 단어가 서로 유사했을 때 K 평균 군집화 알고리즘을 이용해 단어를 확장했다. 즉, 제목의 단어와 내용에 있는 단어가 같은 클러스터링에 포함될 때, 내용에 있는 단어를 자질에 추가하였다. 이를 통해 낚시성 기사 제목 분류 실험에서 제목 기반뿐만 아니라, 확장된 제목으로 실험할 수 있다.

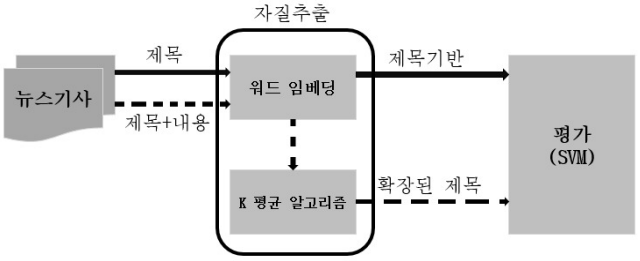
[표 3]은 1,000개의 군집 중에서 한 군집의 예시이다. 주로 '가수'와 연관성 있는 단어들이 같은 군집으로 묶인 것을 알 수 있다.

[표 3] K 평균 군집화 결과

Apink, Apinks, A걸스데이, B1A4, B1A4라, B1A4에, B1A4여성, B1A4와, B1A4음반, B1A4음원, B2ST, B2ST때, B920, BEAST비스트, BEST52현아, BIG8, BJ범프리카, wiki원더걸스, wiki카라아이돌, xi우민은, yayaya, ㅈ됐지, 가면홍백가합전, 가슴치킨가슴, 가식걸, 가요방송, 가요순위, 가요인기가요, 가요팬, 가요프로, 가요프로그램, 가윤, 가은달샤벳가은, 가은달샤벳가은에, 가인, 서현소녀시대서현, 서현염색, 소녀시대, 소녀시대걸그룹, 소녀시대소속사, 소녀시대아이돌, 소녀시대태티서, 태양빅뱅태양, 태연, 태연s2태연, 태태서 등..

2.4 전체 시스템 구조

2.1에서 수집한 기사에서 추출한 기사의 단어들을 기본 자질로 하여 워드임베딩과 군집화 알고리즘을 이용하여 자질을 확장하고, 이를 서포트 벡터 머신(SVM)[6]을 통해 모델링 하였다. 본 논문의 낚시성 기사 제목 분류 시스템의 전체 구성은 [그림 1]과 같다.



[그림 1] 시스템 구조

3. 실험 및 평가

3.1 실험 데이터

실험을 위해 6개의 언론사 홈페이지의 연예 기사를 수집하였다. 수집한 기사들을 2.1절에서 소개한 방법을 통해 낚시성/비낚시성 기사로 분류하였다. 이를 통해 22,729개 신문기사를 수집하였으며, 학습과 평가에 사용된 낚시성/비낚시성 기사는 [표 4]와 같다.

K 평균 군집화 알고리즘에 적용한 단어들은 모두 6,190,821개이며 모두 임베딩되어 있다. 이를 바탕으로 군집의 개수를 200개라고 가정하면 한 군집 당 약 30,000개의 단어가 포함된다.

[표 4] 학습 및 평가 데이터

	낚시	비낚시	전체
학습	500	500	1,000
평가	225	255	450
전체	725	725	1,450

3.2 낚시성 기사 제목 분류 실험

실험은 K 평균 군집화 알고리즘의 K의 범위와 임베딩의 차원을 조정하며 진행하였고, 평가는 서포트 벡터 머신을 사용하여 학습한 모델을 이용해 평가 데이터를 입력하여 낚시성 기사 제목 분류 평가를 진행하였다.

실험은 총 2가지 방법으로 진행하였다. 첫 번째 실험은 제목 기반의 낚시성 기사 제목 분류 실험으로, 각 기사의 제목에 있는 단어들의 임베딩 결과를 각 성분별로 평균을 내어 표현하였고, 자질의 크기는 100부터 1,000으로 늘려가면서 실험하였다.

두 번째 실험은 확장된 제목을 이용한 낚시성 기사 제목 분류 실험으로, 내용에 있는 단어를 더 확장하여 진행하였다. 군집화한 클래스 K개에 대해서 기사의 제목에서 나온 단어와 내용에 나온 단어가 같은 클러스터링에 포함되었을 경우 포함된 모든 단어들을 각 성분별로 평균하였다. 자질의 크기는 100부터 1,000까지이며, 군집의 개수도 100부터 1,000까지 적용하여 실험을 진행하였다.

[표 5] 제목 기반의 낚시 기사 분류 실험 결과

크기	100	200	300	400	500	600	700	800	900	1000
정확도	72.00%	72.00%	77.78%	74.00%	78.67%	77.78%	77.11%	77.11%	**80.00%**	79.78%

[표 6] 확장된 제목을 이용한 낚시 기사 분류 실험 결과

크기 / 군집 개수	100	200	300	400	500	600	700	800	900	1,000
100	79.56%	79.56%	81.78%	81.55%	83.56%	83.33%	82.22%	**83.78%**	**83.78%**	83.33%
200	78.44%	75.33%	79.33%	78.89%	80.44%	80.67%	79.33%	81.33%	78.67%	80.22%
300	73.11%	76.22%	74.00%	74.44%	77.56%	75.78%	76.89%	77.33%	77.33%	78.89%
400	78.22%	79.78%	74.89%	76.00%	79.11%	77.78%	79.33%	78.44%	79.56%	80.44%
500	73.78%	77.33%	78.22%	75.67%	78.22%	76.89%	76.89%	78.67%	78.89%	80.00%
600	73.56%	74.89%	78.89%	78.00%	80.22%	77.78%	80.00%	78.67%	79.33%	80.44%
700	72.89%	76.44%	74.89%	78.00%	78.44%	76.89%	77.11%	77.33%	78.44%	80.22%
800	71.78%	73.33%	75.33%	74.89%	78.67%	76.22%	74.44%	78.00%	77.56%	77.78%
900	74.67%	74.22%	75.11%	73.11%	77.11%	76.67%	75.33%	76.22%	76.89%	77.33%
1,000	73.11%	74.67%	74.89%	74.44%	76.44%	75.11%	76.44%	76.22%	75.56%	77.11%

3.3 낚시성 기사 제목 분류 실험 결과

제목 기반의 낚시성 기사 제목 분류의 결과는 [표 5]와 같다. 자질의 크기가 900일 때 가장 높은 결과를 보였다. 확장된 제목을 이용한 낚시성 기사 제목 분류의 실험 결과는 [표 6]과 같다. 군집의 개수가 100일 때 실험별로 높은 결과를 나타냈으며, 자질의 크기가 클 때 높은 정확도를 나타내었다. 그리고 제목 기반의 낚시 기사 분류실험보다 확장된 제목을 이용한 실험이 상대적으로 더 높은 정확도를 보였다.

이를 통해 본 논문의 서론에서 문제를 제기한 바를 확인한다. 즉, 포털사이트는 독자의 클릭 유도를 위해 낚시성 기사에 많이 사용되는 단어를 제목에 사용함을 확인할 수 있었다. 또한 단어를 확장하는 방법도 효과가 있음을 결과를 통해 알 수 있다.

4. 결론

본 논문에서는 낚시성 기사 제목 분류 시스템을 위해서 기사를 수집했고 이를 수작업으로 판별한 데이터를 가지고 평가실험을 진행했다. 실험의 결과로 제목의 단어들이 낚시성과 비낚시성을 구분해주고 있다는 것을 확인하였다. 즉, 제목의 단어는 자질로서 유효하다. 이를 통해 낚시성 기사 제목에는 낚시를 유도하는 특별한 단어들이 존재함을 알 수 있다. 하지만 여전히 미판단된 기사가 존재하고 이로 인해 신뢰성이 떨어진다. 향후 정확도를 보완하여 처리할 수 있는 방법에 대한 연구를 진행하여야한다.

참고문헌

[1] Kim, Sunjin, “The utilization reality of hooking news titles in portal news service - Focused on major daily newspapers in naver newscast -”, “Journal of digital design”, 2010

[2] Bang Youndduc, “News Selection and Headline Editing of the Internet News”,Korean Journal of Broadcasting and Telecommunication Studies, 2009

[3] Tomas Mikolov 외 3명, “Efficient estimation of word representations in vector space”, 2013

[4] 네이버 뉴스캐스트 옴부즈맨, http://cafe.naver.com/navernewscast

[5] Rand W. M, “Objective Criteria for the Evaluation of Clustering Method”, Journal of the American Statistical Association, 1971

[6] Joachims, Thorsten, “Making large scale SVM learning practical”, “Universität Dortmund”, 1999