워드 임베딩을 활용한 한국어 가짜뉴스 탐지 모델에 관한 연구

심재승[O], 이재준[*], 정이태[*], 안현철(교신저자)[*]
[O]국민대학교 BIT전공,
[*]국민대학교 BIT전공
e-mail: simsoni7@kookmin.ac.kr[O], jrack707@naver.com[*], jeongiitae@kookmin.ac.kr[*], hcahn@kookmin.ac.kr[*]

A Study on Korean Fake news Detection Model Using Word Embedding

Jae-Seung Shim[O], Jaejun Lee[*], Ii Tae Jeong[*], Hyunchul Ahn(corresponding author)[*]
[O]Graduate School of Business IT, Kookmin University,
[*]Graduate School of Business IT, Kookmin University

● 요 약 ●

본 논문에서는 가짜뉴스 탐지 모델에 워드 임베딩 기법을 접목하여 성능을 향상시키는 방법을 제안한다. 기존의 한국어 가짜뉴스 탐지 연구는 희소 표현인 빈도-역문서 빈도(TF-IDF)를 활용한 탐지 모델들이 주를 이루었다. 하지만 이는 가짜뉴스 탐지의 관점에서 뉴스의 언어적 특성을 파악하는 데 한계가 존재하는데, 특히 문맥에서 드러나는 언어적 특성을 구조적으로 반영하지 못한다. 이에 밀집 표현 기반의 워드 임베딩 기법인 Word2vec을 활용한 텍스트 전처리를 통해 문맥 정보까지 반영한 가짜뉴스 탐지 모델을 본 연구의 제안 모델로 생성한 후 TF-IDF 기반의 가짜뉴스 탐지 모델을 비교 모델로 생성하여 두 모델 간의 비교를 통한 성능 검증을 수행하였다. 그 결과 Word2vec 기반의 제안모형이 더욱 우수하였음을 확인하였다.

키워드: 가짜뉴스(fake news), 워드 임베딩(word embedding), 인공 신경망(artificial neural network)

I. Introduction

2016년 미 대선과 영국의 브렉시트 투표와 관련하여 가짜뉴스에 대한 세계적인 관심이 생겨난 이후, 학계에서도 가짜뉴스 탐지 연구에 대한 수요가 증가해왔다. 특히 미국을 중심으로 다양한 가짜뉴스 탐지 연구가 수행되어 왔는데, 한국의 경우 여전히 미국과 비교하여 데이터의 규모와 다양성 측면에서 열세를 보이고 있으며, 가짜뉴스 탐지에 필요한 자연어 처리 기법의 활용에 있어서도 영어에 비해 떨어지는 실정이다. 지금까지 제안된 다수의 한국어 가짜뉴스 탐지 모델들은 모두 빈도-역문서 빈도(TF-IDF)를 통한 탐지모델이었다. 이는 뉴스에 등장하는 단어의 빈도수를 활용하여 문서들 사이의 비슷한 정도를 확인하는 방식이다. 하지만 문장의 문맥 정보까지 반영할 수 없다는 점에서 한계를 갖는다. 텍스트 마이닝의 관점에서 더 진보한 방식으로 여겨지는 밀집 표현 기반의 워드 임베딩의 경우 문장의 문맥 정보를 반영해 주는데, 해외의 경우 다양한 사례가 존재하지만 아직까지 본 연구 주제를 대상으로 하는 국내 사례가 전무하다. 따라서 국내 데이터를 활용한 워드 임베딩 기반의 가짜뉴스 탐지 모형을 연구하는 본 연구에서는 Word2vec 기반의 가짜뉴스 탐지 모형을 제안하고 TF-IDF 기반의 가짜뉴스 탐지모델과의 비교를 통하여 성능 검증을 수행한다.

II. Preliminaries

1. Related works

1.1 가짜뉴스 탐지 연구

지금까지의 가짜뉴스 탐지 연구는 크게 콘텐츠 기반 탐지 연구와 콘텍스트 기반 탐지 연구로 나뉜다(Bondielli & Marcelloni, 2019). 이 중 주된 모델이라 할 수 있는 콘텐츠 기반 탐지 연구는 주로 머신러닝 기반의 분류모델과 결합된 형태로 제안되었으며, 해당 콘텐츠가 가진 언어적 특징과 콘텐츠의 형식에서 나타나는 특징을 이용하는 방식이다. 콘텍스트 기반 탐지 연구에는 가짜뉴스의 배포자나 출처정보를 이용하는 방식과 네트워크 특성을 이용하는 방식이 있다. 국내 뉴스를 대상으로 한 가짜뉴스 탐지 연구들은 서베이 연구를 제외하면 대부분 콘텐츠 기반 연구에 해당하며, 그 가운데에서도

언어적 특징을 이용한 연구에 해당한다. 가짜뉴스의 언어적 특징을 추출하는 방식으로는 TF-IDF가 가장 많이 활용되었으며 정호선(2019)의 경우 처음으로 가짜뉴스 탐지에 감성분석과 구문론적 특징을 활용한 텍스트 분석을 접목했다. 이동호 외.(2018)의 경우 워드 임베딩 방식인 FastText를 활용하였으나, 이는 저자들이 가공한 자체 가짜뉴스 데이터셋을 활용한 연구로서 제목과 본문이 다른 클릭베이트(Clickbait) 형태의 가짜뉴스연구에 해당하여 실제 국내에서 사회적인 물의를 일으켰던 가짜뉴스를 다루는 다른 연구와는 방향이 다르다. 국내 가짜뉴스 탐지 연구에 워드 임베딩 방식이 적용되지 않은 이유는 외국에 비해 데이터셋의 규모가 너무 적고 대부분의 연구에서 활용되는 SNU-Factcheck에서 공개한 데이터의 형태가 짧은 글로 이루어진 가짜뉴스이기 때문에 워드 임베딩의 효력이 십분 발휘되지 못했을 것이라고 추측할 수 있다. 하지만 선행연구들이 발표된 이후로 데이터셋이 조금 더 축적되었으며, Wang(2017)의 경우 짧은 형태의 가짜뉴스 데이터셋에도 word2vec을 적용하여 성공적인 가짜뉴스 탐지 모델을 만든 사례가 있으므로 한국어 뉴스를 대상으로 하는 탐지모델에 워드 임베딩을 활용하는 실험을 해 볼 필요성이 제기된다.

Table 1. Korean fake news detection studies

저자	활용 데이터	특징 추출
박재형 & 김영인(2017)	페이스북에서 공유되는 정치관련뉴스	TF-IDF
윤태욱 & 안현철(2018)	SNU Factcheck	TF-IDF
이동호 외.(2018)	제목과 본문 내용이 일치하는/불일치하는 뉴스 데이터	FastText
현윤진 & 김남규(2019)	SNU Factcheck & Twitter	TF-IDF
정호선(2019)	SNU Factcheck	어휘변수, 구문론 변수, 심리언어학 변수

1.2 워드 임베딩

워드 임베딩이란 단어를 벡터로 나타내는 방법을 통칭하는 표현으로, 크게 희소 표현과 밀집 표현으로 구분된다. 희소 표현은 전통적인 벡터 표현법으로 Bag of word, TF-IDF가 있다. Bag of word는 단어의 순서는 고려하지 않고 출연 빈도만을 가지고 단어를 평가하고, 표현하는 방식이다. 가짜뉴스 탐지 연구를 비롯한 다양한 연구에서 활용되고 있는 TF-IDF는 기본적으로 단어의 빈도수를 가지고 단어를 평가하는데, 여러 문서에서 빈번하게 출현하는 단어의 경우 일반적인 단어로 인식하여 낮게 평가하고, 특정 문서에서 발생하는 단어는 높게 평가하도록 설계되어있다. TF-IDF는 전체 데이터에 나타난 개별 용어의 수만큼 차원을 확장하는 희소행렬을 발생시키기 때문에 과적합의 가능성을 내포하고 있다. 또한, 단어의 문맥 정보와 단어 간 의미의 방향성 등을 반영하지 못하는 한계가 있다.

한편 밀집 표현에 해당하는 방식에는 대표적으로 word2vec이 있다. word2vec(Mikolov et al., 2013)은 딥러닝 기반의 단어 임베딩 기법으로서, 밀집 벡터 표현으로 단어의 언어적 맥락을 재구성한다. 단어를 예측하는 방식은 cbow와 skipgram으로 나뉘며, window 개수, 단어의 최소 빈도수, 프로세스의 개수 등 복수의 하이퍼 파라미터가 존재한다. 본 연구에서는 한국어 가짜뉴스 탐지 문제에 word2vec을 실험적으로 적용해보려 한다.

III. The Proposed Scheme

본 연구의 제안 모형은 word2vec 기반의 가짜뉴스 탐지 모형이다. 제안 모형의 성능을 입증하기 위해 전통적인 텍스트 자질 추출 방법인 TF-IDF 기반의 가짜뉴스 탐지 모형을 비교 모형으로 설정하고, 제안모형과 TF-IDF 기반 모형 간의 탐지 성능을 비교한다. Fig. 1.은 본 연구의 진행 과정을 플로우차트 형태로 표현한 그림이다.

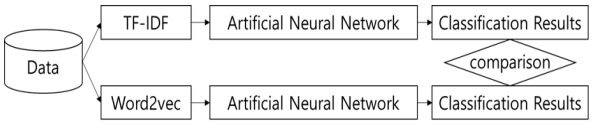

Fig. 1. System Architecture

본 연구는 실제로 국내에서 이슈가 되었던 가짜뉴스를 탐지하는 모형을 만들기 위해, 다수의 선행 연구에서 활용되었던 SNU Factcheck와 네이버 뉴스를 수집하여 2118개의 실험 데이터셋을 구축하였다. 이 중 가짜뉴스는 1059개, 진짜 뉴스는 1059개로 구성되어있다. 가짜뉴스 1059개와 진짜 뉴스 307개는 SNU-Factcheck에서 수집하였으며, 부족한 진짜 뉴스는 네이버 뉴스란에서 2016년부터 2020년까지 올라온 부문별 뉴스 108730개 가운데 무작위로 752개를 수집하여 실험 데이터셋을 완성하였다. 실험을 위한 데이터 분할은 Train 데이터 1440개, Validation 데이터 339개, Test 데이터 339개로 설정했으며 훈련을 위한 특징 추출에는 Test 데이터가 활용되지 않고 이후 훈련된 모델에 의한 벡터 표현에만 사용되도록 하여 좀 더 엄격한 환경을 구축하였다.

다음은 특징 추출 과정이다. 비교모형에서는 TF-IDF가, 제안 모형에서는 Word2vec이 활용되었다. 문장의 전처리는 한국어에 특화된 자연어 처리를 위해 최상혁 외.(2016)의 연구 결과를 참조하여 파이썬 패키지 형태인 Okt 형태소 분석기를 사용하였으며 본 패키지를 통해 TF-IDF의 경우 '명사'를 추출하고, Word2vec의 경우 '형태소'를 추출하였다. 문장의 벡터화 과정에서 TF-IDF의 경우 2118x5034개의 희소행렬이 만들어져 각각의 데이터는 5034개의 차원을 갖게 되었다. 이렇게 데이터의 양에 비해 과도하게 커진 차원을 축소하기 위해 주성분 분석(PCA)을 사용하여 5034차원의 약 10%에 해당하는 50개 차원으로 축소하였다. Word2vec의 경우 PCA와 동일한 환경을 구축하기 위해 dimension을 50으로 고정하였으며, 그 외의 파라미터는 최상혁 외.(2016)의 연구 결과를 참조하여 컨텍스트 윈도우는 5, 단어의 최소 빈도수는 50으로 설정하였으며, 모델은 skip-gram을 활용하였다.

모델링에 활용된 분류기는 ANN(Artificial Neural Network)이며, 모델의 하이퍼 파라미터 튜닝을 통해 제안모형과 비교모형의 분류 성능을 최대화하여 서로 비교하였다. 실험은 sk-learn의 파이썬 패키지를 이용하여 수행하였다. 튜닝 결과 제안모형의 ANN의 경우 1개의 은닉층에 10개의 은닉노드로 구성된 신경망이 구축되었으며 활성화 함수로는 tanh, 최적화 함수로는 lbfgs를 사용하였다. 비교모형의 ANN의 경우 2개의 은닉층에 각각 10개의 은닉노드로 구성된 신경망으로 구축하였으며, 활성화 함수로는 tanh, 최적화 함수로는 adam을 사용하였다. 모든 실험의 random state는 동일하게 고정하였으며 반복 수는 두 모델 다 200으로 설정하여 실험환경을 통제하였다.

IV. Conclusions

제안모형과 비교모형의 성능 비교 지표는 모형의 재현율과 정밀도를 균형 있게 반영하는 지표인 F1 score를 사용하였다. Table 2.는 두 모형의 분류 결과를 비교한 도표다.

Table 2. F1 score Results of models

모형	F1 score
비교모형	훈련 0.792 검증 0.627 예측 0.657
제안모형	훈련 0.793 검증 0.754 예측 0.792

실험 결과 훈련 데이터에서는 두 모형 간에 F1 score 값의 차이가 거의 없었으나 검증 데이터에서 12.7%P, 예측 데이터에서 13.5%P의 차이를 확인하였다.

이어서 실험의 공정성을 확보하기 위해, 정확도(Accuracy)를 평가 지표로 하여 K-fold 교차 검증(K=5)을 수행하였다.

Table 3. Accuracy score Results of models

모형	1fold	2fold	3fold	4fold	5fold
비교 모형	훈련 0.791 검증 0.625 예측 0.655	훈련 0.694 검증 0.588 예측 0.612	훈련 0.653 검증 0.564 예측 0.576	훈련 0.65 검증 0.552 예측 0.635	훈련 0.656 검증 0.607 예측 0.620
제안 모형	훈련 0.790 검증 0.752 예측 0.788	훈련 0.789 검증 0.741 예측 0.777	훈련 0.75 검증 0.754 예측 0.776	훈련 0.778 검증 0.705 예측 0.789	훈련 0.797 검증 0.754 예측 0.714

5가지의 실험 결과를 확인해보면 비교모형의 경우 5fold 평균이 훈련 0.6888, 검증 0.5872, 예측 0.6196이며, 제안모형의 경우 훈련 0.7808, 검증 0.7412, 예측 0.7688로 제안모형의 결과가 월등히 앞서고 있다. 따라서 한국어 가짜뉴스 탐지 연구에 있어서 word2vec을 활용할 경우 TF-IDF보다 뉴스의 언어적 특징을 더 잘 담아낼 수 있음을 확인하였다. 이번 연구는 실제 한국어 뉴스 데이터를 대상으로 워드 임베딩을 적용한 첫 번째 사례라는 데에 의의가 있다. 하지만 언어 기반 탐지 방법론이 갖는, 언어적 특성에 과도하게 의존한다는 단점을 여전히 담고 있다. 따라서 한국어 가짜뉴스 탐지 문제의 경우, SNS 정보와 같은 외부 정보를 결합하면 모델의 성능을 더욱 향상시킬 수 있을 것이다. 또한 분류기로 단순한 형태의 신경망 대신 가짜뉴스를 비롯한 텍스트 분류 문제에서 높은 성능을 보여주는 것이 확인되었던 CNN과 같은 변형모델을 사용할 경우 성능을 한 층 향상시킬 수 있을지, 특징 추출 과정에서 Bert와 같은 Attentional Representations 방식(Luan, et al., 2020)을 적용할 경우 추가 성능 개선이 가능할지에 대한 추가 연구가 필요하다.

REFERENCES

[1] Bondielli, A., & Marcelloni, F., " survey on fake news and rumour detection techniques," information Sciences, Vol. 497, pp. 38-55, Sep 2019.

[2] Choi, S., Seol, J., & Lee, S. G.," On Word Embedding Models and Parameters Optimized for Korean," In Annual Conference on Human and Language Technology, pp. 252-256, Oct 2016

[3] Hyun, Y. and Kim, N., "Text Mining-based Fake News Detection Using News And Social Media Data," Journal of Society for e-Business Studies, Vol. 23, No. 4, March 2019

[4] Jung, H, "Fake News Detectoion Using Content-based Feature Extraction Method," Master Thesis, Feb 2019.

[5] Lee. D-H, et al. "Fake news detection using deep learning" In Proceedings of the Korea Information Processing Society Conference, pp. 384-387, May 2018.

[6] Luan, Y., Eisenstein, J., Toutanova, K., & Collins, M., “Sparse, Dense, and Attentional Representations for Text Retrieval,” arXiv preprint arXiv:2005.00181, 2020.

[7] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. “Distributed representations of words and phrases and their compositionality,” In Advances in neural information processing systems, pp. 3111-3119, 2013.

[8] Park. J-H, and Kim. Y-I, "Development of a Fake News Discrimination System using SVM Classifier," In Proceedings of KIIT conference, Vol. 2017, No. 6, pp. 354-355, June 2017.

[9] Wang, W. Y. " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648.

[10] Yoon. T-U, and H. Ahn, "Fake News Detection for Korean news Using Text Mining and Machine Learning Techniques," Journal of information technology applications & management, Vol. 25, No. 1, pp. 19-32, Mar 2018.