코로나19 가짜뉴스와 진짜뉴스 판별 시스템

이지민*, 이지선O, 우지영*
*순천향대학교 빅데이터공학과,
O순천향대학교 빅데이터공학과
e-mail: {dlwlals7359, 10leegisun@naver.com}, jywoo@sch.ac.kr

COVID_19 fake news and real news discrimination system

Jimin Lee*, Jisun LeeO, Jiyoung Woo*
*Dept. of Bigdata Engineering, Soon Chun Hyang University,
ODept. of Bigdata Engineering, Soon Chun Hyang University

● 요 약 ●

본 논문에서는 코로나19 뉴스와 코로나19 가짜뉴스의 데이터셋을 활용하여 입력 받은 뉴스가 가짜뉴스일 확률을 예측한다. 가짜 뉴스 본문에는 코로나19, 대통령, 정부, 가짜, 언론 등의 키워드의 빈도가 높았다. 위의 키워드를 토대로 나이브 베이즈 모델링을 하여 이를 적용해 가짜 뉴스를 가려내는 웹페이지를 개발하였다.

키워드: 코로나(COVID-19), 가짜 뉴스(fake news), 나이브 베이즈(Naïve Bayes)

I. Introduction

잘못된 정보가 퍼지는 '인포데믹'으로 국민들은 큰 우려와 피해를 받고 있다. 이 상황에서 코로나19와 관련된 뉴스가 신뢰할 수 있는지 알아보는 것은 중요하다. 따라서 이를 해결하는데 도움이 되고 자 '한국언론진흥재단_뉴스 빅데이터_메타데이터_가짜뉴스 '데이터와 '한국언론진흥재단_뉴스 빅데이터_메타데이터_코로나' 데이터를 활용해 가짜뉴스를 판별해주는 웹페이지를 만들고자 한다.

II. Method

공공데이터 포털의 데이터를 기반으로 가짜뉴스의 키워드, 특징, 언론사. 기고자 등을 분석한다. 가짜 뉴스에서 많이 등장하는 키워드를 활용해 코로나19 뉴스 기사를 입력했을 때, 해당 뉴스가 가짜 뉴스인지 진짜 뉴스인지를 예측한다.

III. The Proposed Scheme

1. 데이터 소개

공공데이터 포털의 '한국언론진흥재단_뉴스빅데이터_메타데이터_코로나' 데이터셋과 '한국언론진흥재단_뉴스빅데이터_메타데이터_가짜뉴스' 데이터셋을 활용하였다. 코로나 데이터셋은 코로나 뉴스와 키워드를 확인 가능하며, 일자, 언론사, 기고자, 제목, 키워드, 본문 등을 포함한다.(Fig. 1.) 2020년 1월부터 7월까지 대략 80548개의 데이터로 구축되어 있다. 가짜뉴스 데이터셋은 코로나 가짜뉴스와 키워드를 확인 가능하며, 위의 코로나 데이터와 포함 내용이 같다.(Fig. 2.) 2016년도부터 2020년 12월까지 대략 15911개의 데이터로 구축되어 있다.

2. 데이터 탐색

데이터 분석에 앞서 코로나 뉴스를 많이 다룬 언론사를 비교했다. 코로나 뉴스 중 가짜 뉴스가 가장 많았던 언론사는 YTN, 머니투데이다.(Fig, 3)

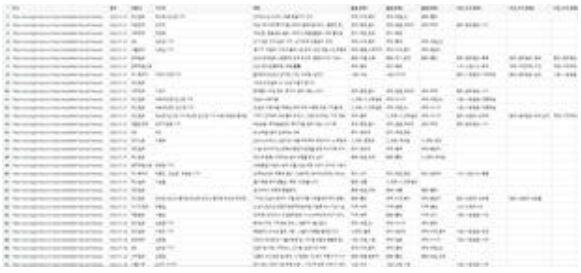

Fig. 1. 코로나 데이터셋

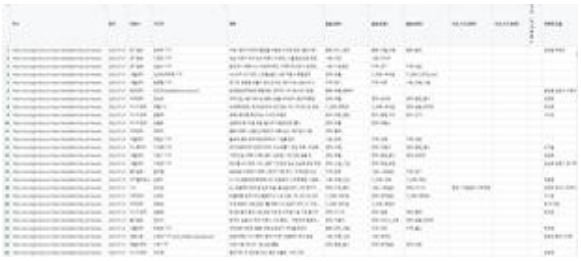

Fig. 2. 가짜뉴스 데이터셋

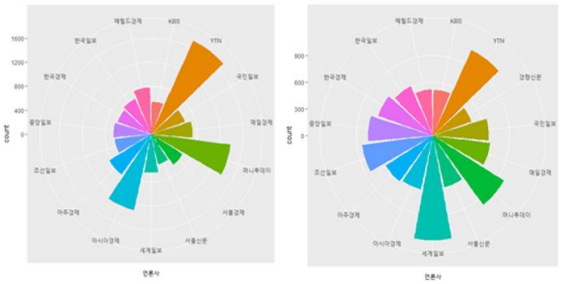

Fig. 3. 코로나 뉴스와 가짜 뉴스의 언론사 비교

3. 데이터 분석

코로나 뉴스 본문에서는 확진자, 접종, 백신, 방역, 지역 등의 키워드가 등장하는 빈도가 높았다. 가짜 뉴스 본문에서는 코로나19, 대통령, 정부, 가짜, 언론 등의 키워드가 등장하는 빈도가 높았다.(Fig. 4.)

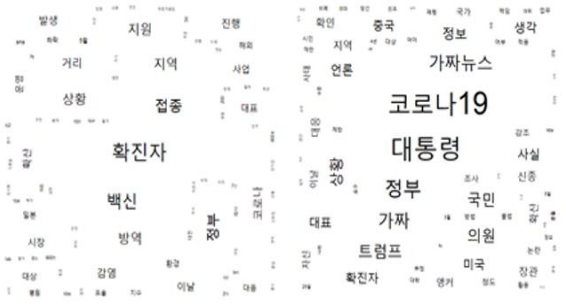

Fig. 4. 코로나 뉴스, 가짜뉴스 키워드 워드클라우드

입력 받은 뉴스가 가짜 뉴스인지 진짜 뉴스인지 나누기 위하여 나이브 베이즈 모델링을 진행했다. 모델링의 결과, 가짜뉴스를 가려내는 정확도는 약 68%로 나타났다.(Fig. 5.)

```
Confusion Matrix and Statistics

          Reference
Prediction 가짜뉴스 진짜뉴스
  가짜뉴스     4635     2777
  진짜뉴스     3680     9022

               Accuracy : 0.679
                 95% CI : (0.6725, 0.6854)
    No Information Rate : 0.5866
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3273

 Mcnemar's Test P-Value : < 2.2e-16
```

	actual		
predicted	가짜뉴스	진짜뉴스	Row Total
가짜뉴스	4635	2777	7412
	805.401	567.583	
	0.557	0.235	
진짜뉴스	3680	9022	12702
	469.976	331.202	
	0.443	0.765	
Column Total	8315	11799	20114
	0.413	0.587	

Fig. 5. 나이브 베이즈

위의 모델을 웹페이지에 적용시켜 웹페이지를 개발하였다. (Fig, 6,)

Fig. 6. 가짜뉴스 판별 웹페이지

IV. Conclusions

이제는 소수가 아닌 다수의 사람들이 저마다의 목적으로 가짜뉴스를 만들기 시작한다. 그리고 이것이 폭발적인 결과를 가져온다. 1인매체의 시대로 접어들면서 '자신이 믿는 바를 말하는' 가짜뉴스의 확산은 더욱 심화될 것이다. 따라서 가짜뉴스를 선별해 내기 위한 시스템을 적용하여 위와 같은 문제를 해결할 수 있다. 또한 웹사이트이기에 접근성이 좋아 보다 많은 활용을 기대해본다.

REFERENCES

[1] 공공데이터포털, 공공데이터포털 (data.go.kr).
[2] 한국언론진흥재단_뉴스빅데이터_메타데이터_코로나, https://www.data.go.kr/data/15069309/fileData.do.
[3] 한국언론진흥재단_뉴스빅데이터_메타데이터_가짜뉴스, https://www.data.go.kr/data/15086437/fileData.do.