인공지능 기반의 언어 생성 모델 분석

이승철, 장용훈, 박창현, 서영석‡
영남대학교 컴퓨터공학과
fatalist316@gmail.com, killerwise@ynu.ac.kr, park@yu.ac.kr, ysseo@yu.ac.kr‡

AI-based language generation model analysis

Seung Cheol Lee, Yonghun Jang, Chang-Hyeon Park, Yeong-Seok Seo‡
Dept. of Computer Engineering, Yeung-nam University

요 약

1989년에 WWW(World Wide Web)이 도입 되면서 세계적으로 인터넷의 보급이 시작되었다. 정보화 시대라고 알려진 3차 산업혁명 이 후로 대량의 정보들이 소셜 미디어를 통하여 생산되었다. 소셜 미디어는 2007년에 인터넷 사용자들 중 56%의 이용률을 보였지만 2008년 2분기에는 75%의 이용률로 증가함에 따라 대부분의 사용자들이 많이 사용하며 의존하게 되었다. 또한 소셜 미디어를 통해 발생 되는 데이터들을 이용하여 기업들은 이윤 창출을 할 수 있다. 하지만 이러한 소셜 미디어는 악의적인 목적을 통해 주가 조작, 정치적 선동 등을 할 수 있는 가짜 뉴스와 허위 정보들을 생성할 수 있으며 이에 따라 대책이 시급하다. 또한 가짜 뉴스는 사람이 글을 작성할 수도 있지만 최근 인공지능 기술의 발달에 따라 프로그램을 통해 자동적으로 생성 될 수도 있다. 본 논문에서는 이와 같은 실제 뉴스와 인공지능을 기반으로 한 뉴스를 분석한다. Kaggle에서 실제 뉴스 데이터를 수집하여 헤드라인을 OpenAI의 GPT-2 언어 모델을 통해 뉴럴 가짜 뉴스를 생성하였다. 파이썬의 NLTK 모듈을 이용하여 전처리를 진행하였고 t-검정과 박스 플롯을 활용하여 분석을 진행하였다. 분석된 주요 속성들을 의사결정트리를 통해 모델 검증을 하였고 k-fold 교차검증을 통해 분류 모델을 평가하였다. 결과로 전체 분류 정확도 평균 89%의 성능을 보여주었다.

1. 서론

정보화 시대라고 알려진 3차 산업 혁명 이래로 대량의 데이터들이 발생하였으며 현 시대에 이르러 국가, 기업, 개인의 커뮤니티 활성을 위한 소셜 미디어의 발전이 중요시 되었다[1]. 인터넷을 이용하는 사용자 중 소셜 미디어를 이용하는 사용자는 2007년에 56%가 사용 하였으며 바로 다음 해인 2008년 2분기에는 75%를 달성하였다[2].

현재 소셜 미디어는 개인 또는 특정 단체의 채널로 운영되며 주로 다양한 사용자들과 정보를 공유하기 위해 사용된다. 공유되는 정보에는 사설 뉴스의 정보도 포함된다. 하지만 전달되는 정보는 진실과 거짓이 존재한다. 거짓 된 정보를 뉴스처럼 전달하는 것을 가짜 뉴스라 하며 Parkinson은 가짜뉴스가 미국의 2016년 대선에 많은 영향을 미쳤다고 언급하였다[3]. 따라서 소셜 미디어에 확산되고 있는 가짜 뉴스의 진위 여부를 판단하는 연구뿐만 아니라 자동화 된 가짜 뉴스의 사실 확인에 대한 연구의 관심도 함께 높아졌다[4]. 과거에 가짜뉴스는 사람에 의해서 작성되었지만 현재에는 인공지능 기술을 기반으로 한 언어 모델을 통해서도 작성이 가능하다. 비영리 인공지능 연구기관인 OpenAI의 GPT-2가 대표적이다. GPT-2는 일부 텍스트의 이전 단어 혹은 문장을 키워드로 제시하면 다음 단어 및 문장들을 예측 할 수 있도록 설계된 언어 모델이다[5]. 인공지능 기반의 언어 모델로부터 작성된 기사를 뉴럴 가짜 뉴스(Neural fake news)라 부르고 있으며 뉴럴 가짜 뉴스는 인간이 작성한 뉴스와 거의 유사하다[6-7].

본 논문에서는 실제 인간이 작성한 뉴스와 인공지능을 기반으로 한 언어 생성 모델이 작성한 뉴스의 분류를 위한 분석을 진행한다.

2장은 OpenAI의 GPT-2에 대한 설명과 관련 연구를 서술하며 3장에서는 분석에 필요한 데이터의 수집 및 전처리 과정을 서술한다. 4장에서는 모델

‡ 교신저자 : 서영석(Yeong-Seok Seo), 컴퓨터공학과(Dept. of Computer Engineering), 영남대학교(Yeung-nam University), Email : ysseo@yu.ac.kr

검증 및 분석 결과를 서술하며 5장은 결론으로 진행한다.

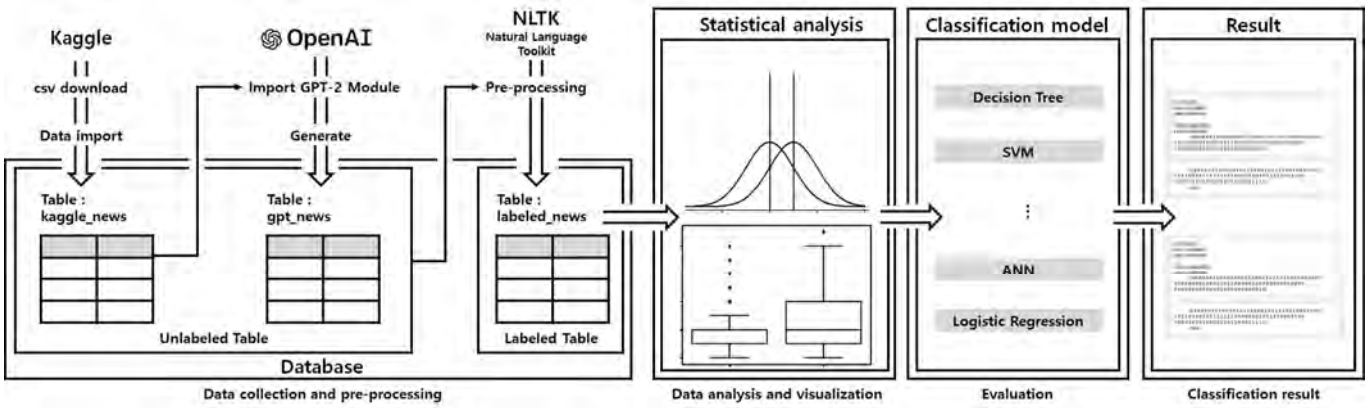
2. 관련 연구

비영리 인공지능 연구기관인 OpenAI는 2019년 2월에 GPT-2를 발표하였다. GPT-2는 텍스트 생성 기능을 가진 인공지능 모델로서 학습에 사용된 데이터는 약 800만 건의 웹페이지 데이터와 약 1,000만 건의 텍스트를 사용하였다. 단어나 문장을 키워드로 제시를 하면 약 40GB의 데이터를 기반으로 한 텍스트를 생성해준다. 또한 생성텍스트의 길이, top-k 등의 매개변수를 설정할 수 있다. 방대한 데이터를 학습한 결과로 GPT-2는 기존에 인간들이 작성하던 뉴스 기사, 블로그 게시글, 소설 등의 글을 작성할 수 있다. 하지만 GPT-2는 특정 목적으로 악용될 소지가 있다. GPT-2는 1시간에 수백에서 수천 개의 뉴스 기사를 생성할 수 있다. 결과적으로 진실과 거짓을 구분 할 수 없는 글들이 소셜 미디어를 통해 전파되어 가짜 뉴스, 주식에 관련한 거짓 정보, 정치적 선동이 가능한 글들로 인하여 소셜 미디어의 사용자들에게 혼란을 야기할 수 있다.

인공지능을 기반으로 한 언어 생성 모델에 관련한 연구는 다음과 같다. Rowan은 텍스트 생성 모델인 Grover와 생성된 텍스트에 대한 탐지 모델을 제안하였다[7]. Joseph는 UN총회에서 발표된 연설 데이터에 대해 사전 훈련된 AWD-LSTM모델을 미세조정하여 정치적 스타일의 텍스트를 생성하는 모델을 제안하였으며 이를 통해 자동화된 텍스트에 대한 위험성과 이를 방지하기 위한 정책의 필요성을 제안하였다[8].

<표 1>실제뉴스와 뉴럴 가짜 뉴스의 유니온을 진행 후 전처리가 진행된 테이블 정의

Table : labeled news		
Column name	Type	Describe
class	int	0 : human, 1 : bot
text	longtext	news text
num_cnt	int	Number of numbers used
word_cnt	int	Number of words used
sentence_cnt	int	Number of sentence used
strong_pos_word_cnt	int	Number of strong positive words
strong_neg_word_cnt	int	Number of strong negative words
weak_pos_word_cnt	int	Number of weak positive words
weak_neg_word_cnt	int	Number of weak negative words
sentimental_score2	double	Positive score 2 = NLTK positive score + NLTK negative score
pos_sentence_cnt	int	Number of positive sentences
neg_sentence_cnt	int	Number of negative sentences
sentimental_score1	double	Emotion score = strong positive-strong negative + (weak positive-weak positive) * 0.5
sentence_score1	double	Sentence score of sentence = positive-negative
sentence_score2	double	Sentence positive score 2 = nltk sentence positive score + nltk sentence negative score
link_cnt	int	Number of links
schar_cnt	int	Number of special characters
noun_cnt	int	Number of nouns
pronoun_cnt	int	Number of pronouns
verb_cnt	int	Number of verbs
modal_cnt	int	Number of auxiliary verbs
adject_cnt	int	Number of adjectives
adverb_cnt	int	Number of adverbs
interject_cnt	int	Number of admirers
interrogat_cnt	int	Number of interrogators
conjuct_cnt	int	Number of conjunctions
unkwon_cnt	int	Number of non-attributes



(그림 1) 데이터 수집과 전처리 및 시각화 과정.

<표 2>각 속성의 T-검정 결과

col_name	p-value
num_cnt	*0.516443
word_cnt	*0.473975
sentence_cnt	*0.25634
strong_pos_word_cnt	*0.646823
strong_neg_word_cnt	**0.035263
weak_pos_word_cnt	*0.804468
weak_neg_word_cnt	*0.723293
sentimental_score2	*0.081138
pos_sentence_cnt	*0.581243
neg_sentence_cnt	**0.001746
sentimental_score1	*0.081138
sentence_score1	*0.127788
sentence_score2	*0.134157
link_cnt	*0.169426
schar_cnt	**1.23E-51
noun_cnt	*0.278986
pronoun_cnt	*0.877456
verb_cnt	**0.018639
modal_cnt	**0.142529
adject_cnt	**0.000252
adverb_cnt	**7.20E-05
interject_cnt	*0.843937
interrogat_cnt	*0.167491
conjuct_cnt	**0.003081
unkwon_cnt	**0.001539

* : p-value >= 0.05 ** : p-value < 0.05

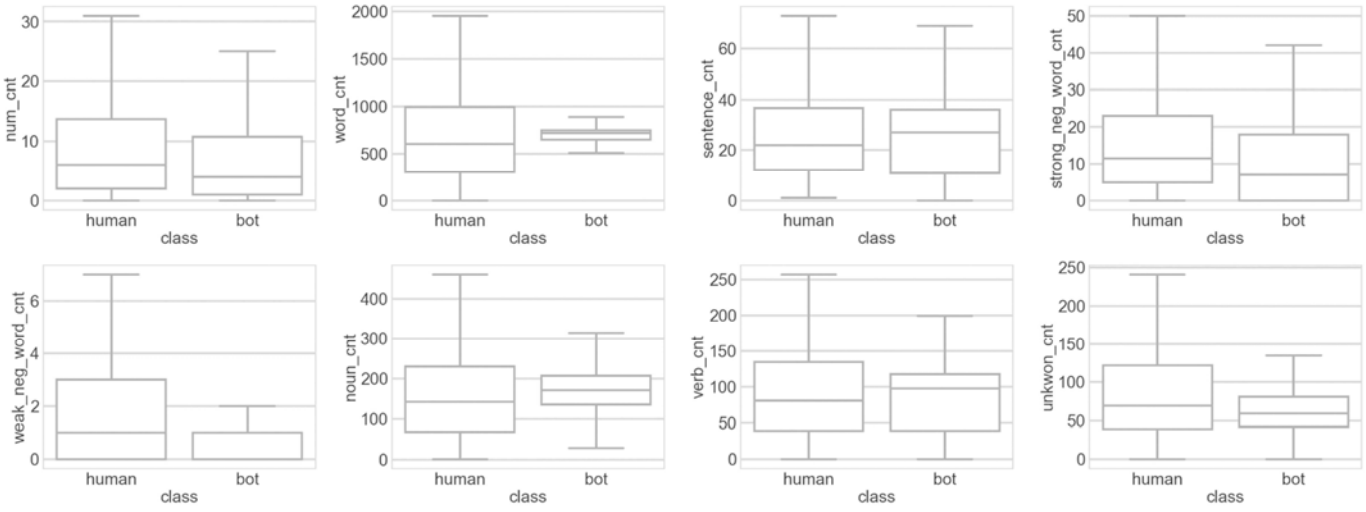
3. 언어 생성 모델을 통한 데이터 수집 및 전처리

본 장에서는 인간이 작성하는 뉴스 기사와 GPT-2를 이용하여 작성된 뉴럴 가짜 뉴스 기사를 분석하기 위한 데이터 수집 및 전처리 과정을 설명한다. 인간이 작성하는 뉴스 기사를 수집하기 위해 Kaggle의 데이터셋을 다운로드 한다. 다운로드한 csv파일을 데이터베이스에 저장하기 위해 데이터 테이블을 정의한다. 그림 1과 같이 GPT-2를 이용하여 실제뉴스의 헤드라인을 키워드로 제시하여 뉴럴 가짜 뉴스를 생성한다. 마찬가지로 GPT-2로 생성한 뉴스의 테이블을 정의하였으며 헤드라인은 뉴럴 가짜 뉴스를 생성하기 위해 제시했었던 키워드로 설정하였다. 실제뉴스 테이블과 뉴럴 가짜 뉴스 테이블을 유니온 한 후 파이썬의 자연언어처리지원모듈인 NLTK를 통하여 뉴스의 내용을 표 1과 같이 25개의 속성으로 전처리 작업을 진행하여 데이터베이스에 저장한다. 최종 레이블 된 테이블을 t-검정과 파이썬의 seaborn 모듈을 이용하여 분석결과를 시각화 하며 분류 모델을 통한 검증을 진행한다.

4. 모델 검증 및 분석 결과

실제 인간이 작성한 뉴스를 실제 뉴스라 통일하며 뉴럴 가짜 뉴스와의 차이를 확인하기 위해 실제 뉴스 300건과 뉴럴 가짜 뉴스 300건을 활용하였다. 표 2는 각 속성에 대해 t-검정을 수행한 결과이며, 그림 2는 각 속성들에 대한 데이터 분포를 박스플롯을 이용하여 유의미한 차이가 있는 속성을 나타낸 것이다. 그림 2의 박스플롯에서 x축은 뉴럴 가짜 뉴스 여부를 의미하고 y축은 각 속성 값의 범위를 나타낸다.

실제 뉴스와 뉴럴 가짜 뉴스간의 유의한 차이를 보이는 속성들은 숫자 사용수(num_cnt), 단어의 수(word_cnt), 문장의 수(sentence_cnt), 강한 부정 단어 사용의 수(strong_pos_word_cnt), 약한 부정 단어 사용의 수(weak_neg_word_cnt), 명사 사용 수(noun_cnt), 미 분류 속성의 수(unkwon_cnt) 등이다. 실제 뉴스와는 달리 뉴럴 가짜 뉴스를 생성하는 GP



(그림 2) 각 속성들의 데이터 분포

<표 3>실제 뉴스와 뉴럴 가짜 뉴스 간
K-Fold 교차 검증 결과

5-fold	accuracy (%)		
	total	bot	human
case 1	91%	98%	85%
case 2	87%	96%	79%
case 3	91%	94%	87%
case 4	92%	98%	88%
case 5	65%	97%	78%
avg	89%	96%	83%

T-2 모듈은 대체적으로 단어의 수가 현저히 적은 것으로 보이기 때문에 사용하는 명사의 수도 적은 것으로 예상된다. 또한 문장의 수가 적은 것으로 보아 인간과는 달리 긴 문장을 구성하는 것이 아닌 여러 짧은 문장으로 구성하는 특성이 있어 보인다. 강한 부정 단어 사용의 수와 미 분류 속성의 수를 제외한 속성들은 t-검정에서 p-value가 0.05보다 큰 값이 계산되었다. 이는 두 속성을 제외한 속성들이 실제 뉴스와 뉴럴 가짜 뉴스간의 차이가 통계적으로 유의함을 의미한다.

실제 뉴스와 뉴럴 가짜 뉴스의 분류를 위해 의사결정트리를 사용하였다. 학습 데이터와 검증 데이터의 균형을 위해 k-fold 교차 검증을 진행 하였으며 결과는 표 3에 나타내었다. 분류 모델의 성능은 k-fold step 4에서 92%의 정확도를 나타내었고 분류 정확도의 평균 89%의 성능을 달성하였다.

5. 결론

미디어의 발전과 함께 소셜 미디어의 확산은 급격하게 증가하며 조직 및 개인의 사설 뉴스가 확산되었다. 사설 뉴스의 진실 및 거짓 여부 판단이 어렵고 인공지능에 기반한 언어 모델을 통하여 뉴스를 작성하게 되면 더더욱 어려워 지게 된다. 본 연구에서는 인간이 작성한 실제 뉴스와 OpenAI에서 개발한 GPT-2 언어 모델을 통해 작성된 뉴럴 가짜 뉴스의 분석 및 모델 검증을 하였다. 실제 뉴스 수집을 위해 Kaggle을 통하여 데이터를 다운로드 하였으며 다운로드 된 데이터의 헤드라인을 GPT-2 언어 모델을 통하여 뉴럴 가짜 뉴스를 생성하였다. 수집한 실제 뉴스와 생성된 뉴럴 가짜 뉴스의 텍스트를 파이썬의 NLTK 모듈을 통해 전처리 작업을 진행 하였으며 이를 분석하기 위해 t-검정과 박스 플롯을 활용하였다. 숫자 사용수, 단어의 수, 문장의 수, 강한 부정 단어 사용의 수, 약한 부정 단어 사용의 수, 명사 사용 수, 미 분류 속성의 수중에서 강한 부정 단어 사용의 수와 미 분류 속성의 수를 제외한 나머지 속성에서 p-value가 0.05보다 큰 것을 통해 통계학적으로 유의미한 것을 확인하였다. 또한 의사결정트리를 통해 분류 모델 검증을 하였으며 k-fold 교차 검증을 통해 데이터의 범위별 분류 정확도의 평균 89%의 성능을 달성하였다.

Acknowledgement

이 성과는 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2017R1C1B5018295). 이 연구는 2019년도 영남대학교 학술연구조성비에 의한 것임

참고문헌

[1] Kyle Hensel, Michael H. Deis, "Using social media to increase advertising and improve marketing.", The Entrepreneurial Executive. Vol. 15, pp. 87, 2010.

[2] Andreas M. Kaplan, Michael Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media.", Business horizons, Vol. 53, No. 1, pp. 59-68, 2010.

[3] Hannah Jane Parkinson, "Click and elect: how fake news helped Donald Trump win a real election.", The Guardian, Vol. 14, 2016.

[4] Jan Christian Blaise Cruz, Julianne Agatha Tan, Charibeth Cheng, "Localization of Fake News Detection via Multitask Transfer Learning.", arXiv preprint arXiv:1910.09295, 2019.

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, "Language models are unsupervised multitask learners.", OpenAI Blog, Vol. 1, No. 8, pp. 9, 2019.

[6] Anne K. Cybenko, George Cybenko, "AI and fake news.", IEEE Intelligent Systems, Vol. 33, No. 5, pp. 1-5, 2018.

[7] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, "Defending against neural fake news.", Advances in Neural Information Processing Systems. pp. 9051-9062, 2019.

[8] Joseph Bullock, Miguel Luengo-Oroz, "Automated Speech Generation from UN General Assembly Statements: Mapping Risks in AI Generated Texts.", arXiv preprint arXiv:1906.01946, 2019.