Detecting Fake News using Machine Learning Algorithms

- 트위터 실시간 데이터 추출->데이터 전처리->다양한 머신러닝 알고리즘

관련 연구

1) SNS 상의 가짜뉴스 탐지(Detecting Fake News in Social Media Networks)

- Clickbait : SNS 상에서 하이퍼 링크 또는 썸네일 링크로 이루어진 거짓 광고, 사용자들의 호기심을 자극하여 클릭을 통해 광고 수익을 높임
- 사용자가 SNS를 사용할 때, 허위 정보가 포함된 사이트를 걸러내기 위함
- Logistic Regression을 사용하여 분류하였으며 실험 결과 99.4%의 정확성을 보임

2) 머신러닝 및 딥러닝 알고리즘을 이용한 가짜뉴스 탐지(Detecting Fake News using Machine Learning and Deep Learning Algorithms)

- 트위터에서 추출한 데이터 중 가짜뉴스를 식별하기 위한 모델 제시
- SVM(Support Vector Machine) / Naïve Bayes Method / Logistic Regression / RNN(Recurrent Neural Network models) 등 잘 알려져 있는 머신러닝 알고리즘 성능 비교
- SVM과 Naïve Bayes가 다른 알고리즘보다 성능이 우수하다는 결과를 보임

3)가짜뉴스 탐지를 위한 머신러닝 패러다임(Which machine learning paradigm for fake news detection?)

- 가짜뉴스 탐지를 위해 다양한 머신러닝 기법 제안
- 8가지 머신러닝 알고리즘을 소개하고 종합적인 성능 평가가 필요하다고 실험 제안

4)N-Gram과 머신러닝 기술을 사용하여 온라인 가짜뉴스 탐지(Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques)

- N-Gram 분석 : 문자열을 N개의 단위로 절단해서 분류하여 빈도를 계산하는 것
- Stop word : '에', '에서', '를', '는' 등 문장에서 불필요한 단어를 뜻함
- BoW : 분해되고 필터링된 단어들을 벡터로 변환하는 것, 글의 특징을 벡터화하여 어떤 글의 카테고리에 속하는지 수학적으로 처리 가능(ex. 많이 등장하는 단어 정보를 벡터로 표현했을 때, 카테고리가 과학이냐 인문학이냐)
- TF-IDF : BoW에서 추출한 특징 단어에서 중요한 거 같은 단어의 가중치를 크게 부여하는 것
- LSVM(Linear Support Vector Machine)을 사용하여 분류하였으며 92%의 정확성을 보임

연구 단계

1) 데이터 검색

- 트위터의 25117개의 트윗 이용

2) 데이터 전처리

- 데이터 전체를 소문자로 변환하고 구두점(반점(,), 온점(.), 물음표(?), 쌍점(:), 쌍반점(;), 줄표(--), 붙임표(-) 등의 부호 제거)
- 가짜뉴스 탐지에 도움이 되지 않는 이모티콘 및 구두점이 사용되기 때문

- 해시태그 제거

3) 데이터 시각화
- 데이터 전체에서 자주 사용된 단어 상위 25개 데이터 시각화
- 데이터를 더 자세히 이해할 수 있도록

4) 토큰화
- 문장이나 단어를 더 작게 나누는 작업

5) 특징 추출
- TF-IDF를 이용하여 각각의 단어에 가중치 부여
- 등장 빈도에 따라 데이터를 요약할 수 있음

6) 머신러닝 알고리즘
- Logistic Regression / Naïve Bayes / LSTM(Long Short-Term Memory) / SVM 등 사용하여 실험

7) 학습 및 테스트

8) 결과
ⓐ Logistic Regression : 93.8%
ⓑ Naïve Bayes : 72.5%
ⓒ LSTM : 50.5%
ⓓ SVM : 92.5%
- Logistic Regression과 SVM의 성능이 우수하다는 결과를 보임

결론
- Logistic Regression과 SVM이 90% 이상의 높은 정확도를 보임
- LSTM은 이미지나 비디오와 같은 구조화되지 않은 데이터에 적합한 것으로 보임
- 이후 이미지나 비디오와 같은 비정형 데이터에서 가짜뉴스를 탐지하는 방향으로 연구를 확장할 것임

TI-CNN: Convolutional Neural Networks for Fake News Detection

- 많은 convolutional layers를 통해 가짜뉴스에 사용된 단어와 이미지에서 특징 추출
- TI-CNN(Text and Image information based Convolutinal Neural Network) 모델 제안
- 고품질 데이터셋을 수집하고 다양한 관점에서 심층 분석
- 텍스트와 이미지 정보 분석에서 CNN을 사용한 통일된 모델을 제안

관련 연구

- 1) 언어적 접근법 2) 네트워크 접근법 두 가지 측면에서 가짜정보 탐지를 실행

1) 언어적 접근법

- 반자동 자연어 처리 기술 사용
- 아마존에 대한 가짜 리뷰에서 감정, 어휘, 스타일 등을 분석하여 다른 가짜 리뷰를 식별
- 단어 분석 기반의 탐지는 가짜를 식별하기에 충분하지 않음

2) 네트워크 접근법

- 네트워크 구조 분석
- 지식 그래프(개체 간의 관계) 분석을 기반으로 가짜 식별에서 61-95%의 정확도를 보임

3) 신경망 접근법

- 딥러닝은 다양한 분야에서 사용되고 있으며, 자연어 처리 또한 딥러닝이 사용될 수 있음
- CNN은 자연어처리 뿐만 아니라 이미지 특징을 추출하는데 효과적임을 알 수 있음

데이터 분석

- 뉴스의 텍스트와 이미지 정보를 수집하여 분석
- 다양한 관점에서 바라본 특징을 분석하여 가짜뉴스를 판단할 수 있는 정보를 찾음

1) 데이터 집합

- 해당 논문에선 20,015개의 뉴스(11,941개의 가짜뉴스와 8,074개의 진짜뉴스)가 데이터셋으로 사용됨
- 가짜뉴스의 경우 240개 이상의 웹사이트에서 얻음
- 진짜뉴스의 경우 뉴욕 타임즈, 워싱턴 포스트 등에서 얻음
- 데이터 집합에는 제목, 본문, 이미지, 작성자 및 웹사이트 등의 정보가 포함
- 진짜와 가짜의 본질적인 차이를 보이기 위해 제목, 텍스트, 이미지 정보만을 사용

2) 텍스트 분석

- 뉴스에 제목이 없으면 제목을 'notitle'로 지정
- 가짜뉴스 제목에 IN, THE, 혹은 의미없는 숫자들이 많음
- 제목에 사용되는 단어들의 빈도 수를 측정한 결과

ⓐ 가짜뉴스 중 상당 수가 제목을 가지지 않음, 소셜네트워크에서 링크로 퍼짐
ⓑ 가짜뉴스는 대문자 수가 많음, 독자들의 관심을 끌기 위함
ⓒ 진짜뉴스는 자세한 설명이 존재함

① 언어 관점

ⓐ 단어 및 문장 수
- 가짜뉴스는 진짜뉴스보다 언어적 표현이 적음
- 진짜뉴스는 평균 4,360개의 단어가, 가짜뉴스는 3,943개의 단어가 존재
- 가짜뉴스끼리도 단어 수 편차가 큰데 이는 가짜뉴스마다 단어가 거의 없거나, 단어가 너무 많다는 것을 의미함
- 문장 또한 진짜뉴스가 더 많은 문장을 가짐
- 진짜뉴스는 84개의 문장이, 가짜뉴스는 69개의 문장이 존재
- 한 문장에서의 단어 수 또한 진짜뉴스가 더 많음
- 진짜뉴스는 언론의 일정한 규칙을 따라 뉴스를 쓰지만, 가짜뉴스는 규칙이 없기에 진짜뉴스는 개수들의 편차가 크지 않지만 가짜뉴스는 범위가 넓음

ⓑ 물음표, 느낌표 및 대문자
- 진짜뉴스는 가짜뉴스보다 물음표 수가 적음
- 질문이 많은 뉴스는 감정을 격화시킴
- 진짜뉴스와 가짜뉴스 모두 감탄사는 거의 사용되지 않음
- 그러나 가짜뉴스는 명령어를 사용하여 감정을 격화시키는 경향이 있음
- 가짜뉴스는 진짜뉴스보다 대문자가 훨씬 많음
- 독자의 관심을 끌고 믿게 하기 위함

ⓒ 인지적 관점
- 부정의 의미를 지닌 배타적 단어(‘but’, ‘without’, ‘however’), 부정적 단어('no', 'not')등 여부 조사
- 가짜뉴스보다 진짜뉴스가 부정을 더 자주 사용함
- 배타적/부정적 단어를 사용하려면 내용이 더 구체적이고 정밀해야함
- 가짜뉴스를 부정을 할 때 사용할만한 근거/설명이 부족함

② 심리 관점
- 뉴스에서의 1인칭 대명사 사용 조사
- 가짜뉴스는 '우리'와 '나' 등 자신에 대한 언급을 최소화하는 언어를 사용함
- '나는 책을 훔치지 않았다.'라고 표현하면 될 말을 '도둑질은 올바른 행동이 아니다'로 돌려 말하는 등 정확한 내용을 언급하지 않음
- 내용의 깊은 부분을 언급하는 것을 피함

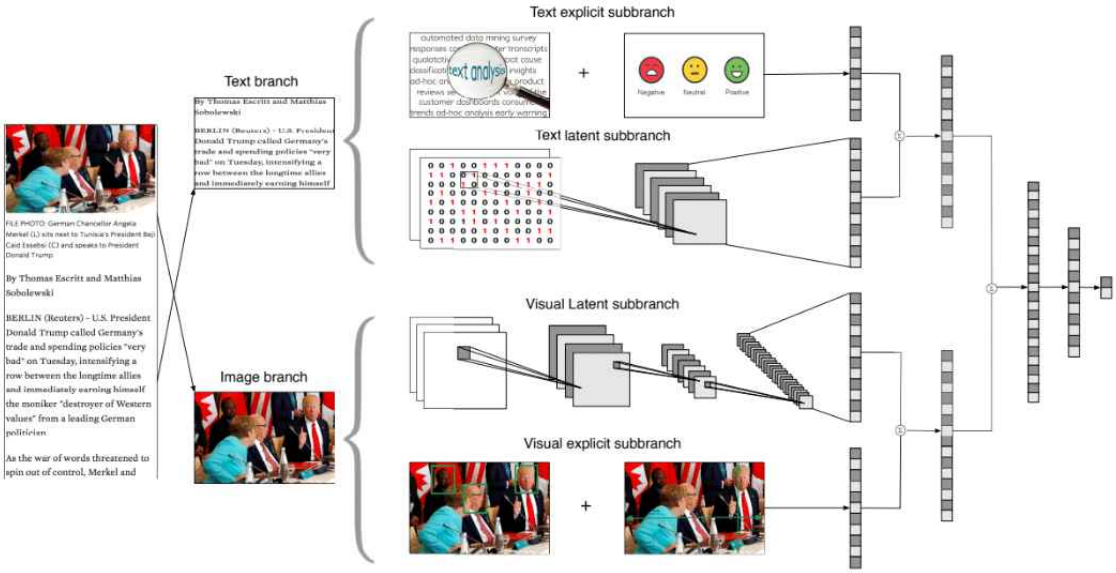
③ 어휘적 다양성
- 진짜뉴스는 가짜뉴스보다 다양한 어휘적 표현이 사용됨

④ 감정 분석
- 진짜뉴스는 가짜뉴스보다 긍정적임
- 가짜뉴스는 사람을 속인다는 죄책감 때문에 내용에서 부정적인 감정이 보일 수 있음

3) 이미지 분석

- 뉴스에 나오는 이미지 특징 분석
- 가짜뉴스보다 진짜뉴스에 인물 사진이 많다는 것을 알 수 있음
- 가짜뉴스는 동물이나, 풍경처럼 관련 없는 이미지를 많이 가지고 있음
- 진짜뉴스는 가짜뉴스보다 해상도가 더 높음

TI-CNN 모델 구조



- 텍스트와 이미지에서 각각 특징을 추출하기 위해 두 개의 병렬 CNN 사용
- 각각 추출 후 텍스트와 이미지 특징을 융합하여 가짜뉴스 탐지
- 텍스트 CNN에는 Sigmoid
- 이미지 CNN에서의 Sigmoid 사용은 Gradient Vanishing(Gradient Backpropagation시 미분값 소실) 현상 발생 가능성이 높아 ReLU 사용

실험

- 데이터셋의 80%를 트레이닝 데이터로 사용하고, 10% 데이터를 검증, 나머지 10% 데이터를 테스트에 사용함

실험 결과

- 논문의 모델과 다른 방법들을 비교함

Method	Precision	Recall	F1-measure
CNN-image	0.5387	0.4215	0.4729
LR-text-1000	0.5703	0.4114	0.4780
CNN-text-1000	0.8722	0.9079	0.8897
LSTM-text-400	0.9146	0.8704	0.8920
GRU-text-400	0.8875	0.8643	0.8758
TI-CNN-1000	**0.9220**	**0.9277**	**0.9210**

- 실험 결과 이미지 정보만으로는 가짜뉴스를 판단하기 어려움
- GPU와 LSTM은 긴 시퀀스에 비효율적임
- TI-CNN이 다른 방법들보다 성능이 높은 것으로 보임

결론 및 향후 작업

- 텍스트와 이미지 정보를 결합할 수 있는 통합 모델 TI-CNN 제안
- 향후 다른 언어로도 가짜뉴스를 판별할 수 있도록 프랑스 자료를 수집할 것
- GAN 또한 텍스트와 이미지를 평가하는 방법이 될 것임