Fake News Detection: A long way to go

※ 가짜뉴스의 정의, 특성, 식별에 대해 설명하고 가짜뉴스 식별판별 알고리즘 및 데이터셋에 대해 소개

1. 가짜뉴스에 대한 정의

- 부분 vs 전체
  - 부분적 가짜뉴스 : 중요한 뉴스의 일부가 누락되고, 단편적인 정보만 나타남
  - 가짜 뉴스 : 일부 또는 완전히 새로운 것을 제공하기위해 제작된 전체 뉴스
- 가짜 뉴스의 종류
  1) 풍자 : 합법적인 뉴스기사로 속이는 경우
  2) 선전 : 제작물을 포함하고 합법적인 뉴스기사로 포장 될 때
  3) 오해의 소지가 있거나 내용 이외의 정보 : 제작 지원을 제공 할 때
  4) ClickBait : 제작물을 포함하고 합벅적인 뉴스 기사로 패키지 될 때
  5) 음모론 : 합법적 인 뉴스 기사로 패키지 된 경우
  6) 사기 : 유머 또는 악의적 인 속임수.

2. 가짜뉴스의 특성

1) 낯선 웹 사이트에서 온 것.
2) URL이 이상하거나 뉴스 매체와 일치하지 않습니다.
3) 광고 제목이 터무니 없거나 기사와 일치하지 않습니다.
4) 날짜가 오래되었거나 보장 범위가 없을 수 있습니다.
5) 저자를 나열하지 않아 조사 할 수 없습니다.
6) 청구 증명을 제공하지 못하는 것.

3. 가짜뉴스의 식별

A. 수동적 가짜 뉴스 식별(인간이 판단)

1) 뉴스는 무작위가 아닌 정통 출처에서 온 것입니다.
2) 헤드 라인이 문제와 관련이 있는지 여부. 물음표의 차이조차도 과감한 오해를 일으킬 수 있습니다
3) 기사 작성자가 신뢰할 수 있는지 확인하십시오.
4) 정보의 주장을 뒷받침하는 출처는 다수의 확실한 출처에 대한 동일한 정보의 가용성과 함께 유효해야합니다.
5) 풍자, 냉담한 의견 또는 농담의 형태의 뉴스가 전혀 없다는 것을 잘 알고 있어야합니다
6) 논리와 신념이 뉴스의 판단과 모순되는지 여부

B. 보다 자동화된 식별

1) N-gram
2) 문자기반 또는 단어기반 정보검색
3) 가짜뉴스를 지원 하기 위한 제작되는 이미지

C. 기타 식별 방법

1) 첫 번째 요점은 신뢰할 수 있고 검증 된 일부 뉴스 사이트의 기존 뉴스를 비교하여 확인할 수 있습니다.
2) 간단한 질문이 수정 된 문장 부호만으로 잘 알려진 문장으로 바뀔 수 있습니다.
   ex)오사마 빈라덴은 죽었다? => 오사마 빈라덴은 죽었다 ! [의미변경]

4. 가짜뉴스 판별 알고리즘

1) 딥러닝

저자명 : Gaurav Bhatt Bhatt

논문명 : [Combining Neural, Statistical and External Features for Fake News Stance Identification]

사용기술 : 심층반복모델, 신경가중 N-GRAM, Bag-of-worlds Model 사용

저자명 : Sarah A. Alkhodair et al

논문명 : [Detecting breaking news rumors of emerging topics in social media]

사용기술 : 데이터마이닝, 비지도학습 World2vec 모델, 반복 신경망 모델 사용

저자명 : Aswini Thota

논문명 : [Fake News Detection: A Deep Learning Approach]

사용기술 : 딥러닝아키텍처를 사용, FNC-1 데이터 셋 사용.하지만 문장부호 및 주요단어 고려 X

2) 머신러닝

저자명 : Georgios Gravanis et al.

논문명 : [Behind the Cues: A benchmarking study for Fake News Detection, Expert Systems With Applications]

사용기술 : 컨텐츠 기반 기능 , 머신러닝 알고리즘 , Adaboost , Bagging, UNB 데이터셋 앙상블 알고리즘 + SVM 데이터셋 사용, 가짜뉴스를 분류하는데 95% 신뢰성 , 단 기사의 진위여부는 알지 못함

3) 기타

저자명 : SD Samantaray &Geetika Jodhani

논문명 : [Survey on Automated System for Fake News Detection using NLP & Machine Learning Approach]

사용기술 : 세개의 텍스트셋을 사용하여 텍스트 유사성 기능 및 NLP, N-gram(문자기반 유사성), TF*IDF(용어기반유사성),Cosine 유사성(말뭉치 유사성), 기사의 진위여부 O

저자명 : Chaowei Zhang et al.

논문명 : [Detecting fake news for reducing misinformation risks using analytics approach]

사용기술: 진위 여부 구별, 가짜뉴스 유형 분류

저자명 : Eugenio Tacchini et al

논문명 : [Some Like it Hoax: Automated Fake News Detection in Social Networks]

사용기술 : Logistic regression, 클라우드 소싱 알고리즘, FaceBook 소셜미디어에 관련

기준	딥러닝	머신러닝	기타
개념	RNN 인공지능	SVM,AdaBoost,Bagging	N-grams,character,world similarities/analytics/cloudsourcing algorithms
기반	Context-Based	Context-Based	Content -Based
페이크뉴스 유형 타겟	Stance related, rumour	풍자,가짜,가짜뉴스게시	의견 기사, 사기

[가짜 뉴스의 3 가지 주요 유형의 도표]

5. 가짜뉴스 탐지 데이터 SET

* 가짜뉴스 탐지 데이터셋 특성

1. 데이터 세트에 실제 뉴스가 있어야 도구가 뉴스를 거부하지 않습니다.
2. 기본 목표가 도구에 의해 받고 있는지 확인하는 가짜 뉴스가 있어야합니다.(...?)
3. 데이터 세트에는 가짜 뉴스 전체에 대한 다양한 뉴스가 주관적으로 포함되어야합니다.
4. 뉴스에서 가짜 또는 실제 뉴스의 백분율을 정의하는 기준.
5. 데이터 세트가 제공된 가짜 뉴스의 실제 버전을 포함합니다.

* 가짜뉴스 탐지 데이터셋 종류

1. Kaggle : 가짜 뉴스 탐지를 위한 충분한 뉴스 기사로 깔끔한 구조화 된 정보를 가지지만 뉴스가 부족하고 부분 뉴스와 함께 실제 뉴스 기사를 포함하지 않습니다. Kaggle의 가짜 뉴스 데이터 세트는 "가짜 뉴스"소스 목록을 사용하는 BSDetector 도구를 기반으로 합니다. “가짜 뉴스”출처 목록을 사용한다
2. PolitiFact.com : 웹 사이트를 통해 얻은 짧은 정치적 진술을 포함합니다. 각 진술에는 저자, 상황, 진실성 레이블 및 그러한 레이블에 대한 정당성이 주석으로 표시되어 있습니다. 뉴스에 필요한 현실의 스펙트럼을 포함합니다. 진부함 등급에 대해 6 가지의 세분화 된 레이블을 가짐 : 새빨간 거짓말, 거짓, 간신히 사실, 반 참, 대부분 참, 참. 이 데이터 세트는 자동 가짜 뉴스 감지에 사용될 수 있습니다. 이 모음은 자세 분류, 인수 마이닝, 주제 모델링, 소문 감지 및 정치적 NLP에 사용할 수 있습니다. LR과 같은 텍스트 분류 모델과 함께 연구및 SVM.
3. Fever : Thorne et al는 Wikipedia 문장을 변경 한 다음 Wikipedia 기사에서 그러한 주장에 대한 증거를 제공함으로써 자신이 작성한 진술을 포함했습니다.
4. Emergent : Ferreira와 Vlachos는 소문이 제기 된 주장과 관련 뉴스 기사를 포함하는 데이터 세트를 제공하여 정확성을 설명합니다. 그러한 데이터 세트의 목적은 타겟 분류이다.
5. CREDBANK : CREDBANK는 6 천만 트윗을 포함하는 대규모 데이터 세트입니다. 이 데이터 세트는 위조트윗을 식별하기 위해 트위터로 전송됩니다. 트윗은 주제 모델링을 통해 이벤트로 그룹화됩니다.
6. George McIntire : 그의 데이터 세트에는 뉴스의 현실성을 위해 스펙트럼의 양면을 사용하여 도구를 테스트하기위한 많은 뉴스탐색이 포함되어 있습니다. 그러나 뉴스 카테고리가 100 % 옳지 않을 수도 있고 뉴스와 뉴스 사이에 관련이 없다는 사실을 고려하면 그다지 효율적이지 않을 수도 있습니다.
7. FakeNewsNet : 뉴스 컨텐츠, 소셜 컨텍스트 및 공간적 시간 정보가 포함 된 데이터 저장소입니다. 가짜뉴스 감지 문제에 매료되어 전 세계 연구원을 언급하는 뉴스.
8. UNBIASED : 다양한 뉴스 소스를 통합하고 몇 가지를 충족시키는 "UNBiased"(UNB) 데이터 세트 치우친 결과를 피하기위한 표준 및 규칙. 분류 작업
9. FNC-1 Challenge : FNC-I 챌린지에 사용되는 Emergent의 특정 데이터 세트는 소문 디버킹을위한 디지털 저널리즘 프로젝트입니다. 데이터 세트에는 뉴스 기사의 본문, 뉴스 기사의 헤드 라인 및 기사와 헤드 라인의 관련성 (자세) 레이블이 포함됩니다.