- 논문 초안

- 표절을 피하고 객관적인 증명자료 (백데이터) 에대한 참고문헌을 맨마막 하단에 첨부할것.

-

1.서론

가짜뉴스는 분야를 막론하고 한국을 비롯하여 전세계에서 주목을 받고 있는 실정이다.

이러한 가짜뉴스는 대중들이 접하는 일반적인 활자매체(신문, 잡지, 서적, 포스터)뿐만아니라 영상매체(TV), 음성매체(라디오), 디지털미디어(인터넷, SNS, 스마트폰)를 통하여 정치적이 거나 경제적 목적을 가지고 가짜뉴스가 유포되고 있으며, 이로 인해 개인을 비롯하여 기업과 국가가 막대한 심리적/경제적 비용을 낭비하고 있으며 사회적 문제로 대두 되고 있다.

가짜뉴스는 허위정보, 오인정보, 거짓정보, 패러디/풍자적, 루머/유언비어의 유형으로 나뉘어 지며 각각의 유형에 따른 가짜뉴스를 접하게되는 대중들은 개인이 가지고 있던 기존 정보에 대한 견해에 대한 변화와 사실이 아님을 인식하지 못한 정보의 습득 및 진실을 가장하여 고의로 조작된 정보들을 수용 할 수 도 있으며, 근거없이 퍼지는 소문 등의 전파로 정보의 불확실성이 SNS나 포털사이트 를 통해 공유되기도 한다. 이러한 가짜뉴스의 사회적 문제를 진단하고 가짜뉴스로 인한 피해를 줄이기 위해 기업들과 정부에서는 가짜뉴스를 탐지하기 위한 관련 연구 및 대응방안을 찾기위한 시스템 개발에 대해 활발한 논의가 이루어지고 있다.

그러나 실제뉴스와 가짜뉴스를 탐지해내는 것은 각자 개개인의 서로 다른 주관적인 견해가 반영 됨으로 가짜뉴스 탐지하는데에는 많은 어려움이 있으며, 이를 해결 하기위해 전문가들이 사실을 근거로된 논리적인규칙으로 검증해내어야 한다. 가짜뉴스 탐지 기법의 대표적인방법으로는 기존의 전문가 기반탐지, 집단지성 기반 탐지 등 비기술적 접근과, 인공지능 기반 탐지, 시맨틱 기반탐지등 기술적 접근으로 나뉘어 진다. 비기술적 접근방법은 검증 과정 및 결과가 명확하고 공신력이 있다는 장점을 가지고 있으나 소수 전문가의 지식에 전적으로 의존함으로써 정보에 대한 전수조사가 어렵고 상대적으로 많은 분석 시간이 요구된다는 단점을 가진다. 집단지성 기반 탐지의 경우에도 대중들의 적극적인 참여가 전제되어야 한다는 것과 검증 결과에 대한 신뢰성을 보장할 수 없는 단점을 가진다.

한편 기술적 접근의 경우 비기술적 접근의 단점을 보완 할 수 있지만, 많은 양의 기계 학습 데이터가 요구되거나 관련된 지식 데이터 축적을위해 상당한 비용과 노력이 요구된다는 한계를 가진고 있다.

* 아래에 본문이 있습니다.

2.조사 및 연구

가짜뉴스 탐지를 하기 위해서 국외에서는 Kaggle, BuzzFeed, Liar, Emergent, PolitiFact, Snope등의 사이트에서 가짜뉴스 데이터셋 수집하고 만들어 기계학습과 텍스트마이닝을 활용하여 다양하고 활발한 연구가 수행되고 있는 반면 국내에서는 SnuFactCheck,Kinds 등의 사이트가 운용되고 있지만 아직 가짜뉴스 연구에 관련된 논문의 수가 미비하고 가짜뉴스 탐지를 하기위한 한글 데이터셋에 대한 정의와 한글 가짜뉴스 데이터를 수집하는데 어려운 점이 있다. 이를 조금이나마 해결하기 위해서 본 논문에서는 가짜뉴스 탐지하기 위한 한국 데이터셋을 정의하고 한글 가짜뉴스 탐지데이터를 수집하기 위해 현재 사용되고 있는 국외의 가짜뉴스 탐지 데이터셋기준에 근거하여 메타데이터를 선택하여 한글 가짜뉴스 탐지 데이터 수집한다.



다음은 영/한국어 가짜뉴스 메타데이터에 대해 조사해보았다.

첫번째 테이블은 가짜뉴스 데이터셋을 가지고 있는 사이트와 그 사이트에 대한 메타데이터별로 정리를 한 부분이다.

영어 가짜뉴스 데이터셋 및 사이트는 Kaggle,BuzzFeed,Liar,Emergent,PolitiFact,Snope 6가지로 분류되고, 한국어 가짜뉴스 데이터셋 및 사이트는 SNU FactCheck 1개로 분류 한다.

표의 맨 왼쪽은 영어/한국어의 가짜뉴스 메타데이터 요소를 나타내고 있으며, 가짜뉴스 탐지로 자주 사용되는 데이터셋을 가지는 사이트에 따라 비교를 나타내었다. 영어/한국어 가짜뉴스 메타데이터에서 사용되는 요소 중 사실유무 라벨, 제목, 본문은 모든 사이트에서 사용된다. 뉴스 데이터의 개수마다 번호가 1씩 늘어나는 번호요소는 Emergent 사이트를 제외하고 사용된다. 뉴스의 기사날짜는 Liar 사이트를 제외하고 사용 된다. 뉴스의 종류는 Liar와 SNU FactCheck 사이트를 제외하고 사용 된다. 뉴스의 기사의 주장을 나타내는 요소는 영어 사이트에서는 모두 사용되지만 한글 사이트에서는 제외 된다. 뉴스 기사의 원문 출처를 타나내는 기사원본주소는 Liar와 SNU FactCheck를 제외하고 사용 된다. 뉴스 기사의 주장을 인용하는 요소 및 영어/한국어 뉴스 사이트상 기사제목,본문,날짜,저자는 BuzzeFeed, Liar, SNU FactCheck를 제외하고 사용 된다. 뉴스 사이트상에서 가짜뉴스를 확인 할 수있는 사이트상 주소는 Liar와 SNU FactCheck를 제외하고 사용 된다. 뉴스 출처 및 요약 요소는 Kaggle과 Emergent에서 사용 된다. 뉴스를 대표하는 사진의 요소는 Kaggle과 BuzzFeed에서 사용된다. 뉴스의 기자명 요소는 Liar와 SNU FactCheck를 제외하고 사용된다. 각 뉴스를 가져온 공식홈페이지의 요소는 Kaggle, BuzzFeed, SNU FactCheck에서 사용 된다. 뉴스를 연설한 연설자, 연설자의 직업, 연설자의 당소속을 나타내는 요소는 Kaggle, Liar, SNU FactCheck에서 사용 된다.

두번째 테이블은 미디어별로 특성을 가지는 메타데이터들에 대한 정리를 해보았다.



모든 미디어들은 제목, 날짜, 출처, 기자명과 각각의 매체 특성에 따른 뉴스본문 내용에 따른 메타데이터를 구성하고 있다.

활자매체는 사진의 주소, 뉴스의 종류, 뉴스에 대한 요약, 뉴스기사의 주장에 대한 메타데이터를 추가적으로 가지고 있으며, 영상매체는 뉴스본문내용을 알 수 있는 동영상 Url과 뉴스 요약에 대한 메타데이터를 추가적으로 가지고 있다. 음성매체는 영상매체와 마찬가지로 음성 뉴스에 대한 본문을 알 수 있는 음성 Url을 메타데이터를 추가적으로 가지고 있다. 뉴미디어는 웹포털사이트와 SNS와 분류를 할 수 있는데, 웹포털 사이트는 위와 같이 SNS관련 정보의 메타데이터를 제외하고 모두 포괄하고 있으며 동영상과 사진 Url은 뉴스에 대한 부연설명에 따른 메타데이터로 구성되어 있다.. 포털사이트와 관련된 정보로는 포털사이트상의 기사의 제목/본문/날짜/저자/주소를 가지고 있으며, SNS는 뉴스종류와 포털사이트 관련정보의 메타데이터를 제외하고 모두 포함 되며 웹포털 사이트와 마찬가지로 동영상과 사진 Url은 뉴스에 대한 부연설명에 따른 메타데이터로 구성되고 있다. 추가적으로 SNS 관련정보 메타데이터는 SNS상의 ID와 SNS 댓글, 추천/비추천의 메타데이터가 구성 되고 있다.

3. 결과 및 제안

우리는 국외와 국내의 가짜뉴스 데이터셋에 대해 조사를 하였고, 미디어별 메타데이터에 대한 조사를 다음과 같이 하였다.

다양한 메타데이터중 우리가 제안하는 가짜뉴스 탐지를 위한 메타데이터 다음과 같이 정의하고자 한다. 첫번째로 이미 국외 및 국내에서 구성된 가짜뉴스 데이터셋을 가지고 있는 사이트의 메타데이터중 모두 포괄하는 메타데이터 사실유무라벨, 제목, 본문 3가지와 두번째로 미디어별로 분류한 메타데이터들중 모든 미디어들이 포함되는 메타데이터인 제목, 날짜, 출처, 기자명 4가지의 메타데이터를 합처 사실유무라벨, 제목, 본문, 날짜, 출처, 기자명

6가지의 메타데이터로 정의를 하고자 하며, 미디어 특성별로 메타데이터를 구성해 본 결과 활자매체, 영상매체, 음성매체는 메타데이터 사용이 뉴미디어의 메타데이터보다 비교적 적으며, 대중들이 접하는 대다수의 가짜뉴스들이 웹포털 사이트 및 SNS를 통해 아주 빠르게 확산되고 있어 우리가 제안하고자하는 가짜뉴스 탐지 메타데이터는 뉴미디어기반의 사용되는 메타데이터에 대해 좀 더 연구를 하여야 할 것 같다고 생각 된다.

* 외국은 다양한 데이터 (설명 자료 추가) 국내는 미비한 데이터 ( △)
* 어떤 미디어든 상관없이 공통으로 들어가는 메타데이터 정의. (o)
* 미디어별로 새롭게 추가 되는 메타데이터 (o)
* 공통 메타데이터 테이블 만들기, 미디어별 메타데이터 테이블 만들기.(o)
* 다양한 메타데이터 중 우리가 제안하는 가짜뉴스 탐지를 위한 메타데이터는 (ex.title, text, ….) 정의 (o)
* AS - IS => TO - BE To be 에대한 근거 작성 (o)
* 우리만의 메타데이터 정의 및 제안 (o)
* 글에서 제안하는 메타데이터 제시할때, 텍스트 기반으로 할 것인지, (o)

-- 조금더 살들을 덧붙여야할것같습니다. 그리고 교수님이 주실 보고서형식에 맞춰서 바꾸어야할것 같습니다.

활자매체 : title/text/date/author/source(출처)

영상매체 : title/text/tv\_url/author/date/

음성매체 title/radio\_address/author/

뉴미디어 SNS와 웹사이트 포털 : 주장 claim/제목title/날짜date/내용text/동영상url/사진img/sns\_id(tweet,facebook)/sns\_comment(tweet, facebook)/ recommand