유해 해시태그 비율 기반의 유해 정보 판단 및 수집 시스템

장정현○ 나스리디노프 아지즈
충북대학교 소프트웨어학과
sky456139@chungbuk.ac.kr, aziz@chungbuk.ac.kr

Harmful hash tag ratio based harmful information judgement and collection system

Jeong-Hyeon Chang○ Nasridinov Aziz
Dept. of Computer Science, Chung-Buk National University

요 약

최근 SNS(Social Network Service)를 이용하는 사용자의 수가 증가하면서, SNS상에는 여러 유형의 정보들이 대량으로 사용자들에게 제공 되고 있다. 이러한 여러 유형의 정보들 중 최근에는 불법 또는 유해 정보를 SNS를 통해 공유하는 등 SNS를 악용하는 사례가 사회적 문제로 대두되고 있다. 본 논문은 이러한 문제를 해결하기 위해 다음과 같은 해시태그 비율 기반의 유해 게시물 판단 기법을 제안한다. 해당 기법은 SNS상의 실제 유해게시물들로부터 해시태그 용어를 수집하여 유해 단어 데이터베이스를 구축하고, 유해 용어를 통해 실시간으로 SNS 게시물을 수집한다. 이후 수집된 게시물내의 해시태그 용어를 추출하고, 유해 단어 데이터베이스를 사용하여 용어의 유해 여부를 확인, 유해 용어가 사용된 해시태그의 비율이 과반수를 초과하는 경우 해당 게시물을 유해하다고 판단한다.

1. 서 론

최근 SNS(Social Network Service)를 이용하는 사용자의 수가 증가함에 따라 SNS상에서 다양한 유형을 가지는 대량의 정보들이 공유되고 있다. SNS에서 공유되고 있는 정보들은 사용자들이 생산해낸 정보들로써, 정보 전달의 목적 또는 개인 사생활에 대한 정보 등 다양한 유형을 가지고 있으며, 이들 정보는 SNS을 사용하고 있는 다른 사용자들에게 제공되게 된다. 이러한 SNS의 특징은 많은 사용자들은 자신의 글이나 사진, 영상 등을 친구 혹은 타인과 공유를 할 수 있고, 사회적 관계망을 형성할 수 있다는 순기능을 지니고 있다.

하지만 이와 반대로 SNS상에서 많은 사용자를 대상으로 불법 또는 유해 정보 가 공유되는 등 SNS의 역기능이 사회적 문제로써 인식되고 있다.

이와 관련된 연구 사례인 '소셜 미디어를 통한 불법·유해정보 유통 실태 및 대응 현황: 국내사례' [1]에서 조사한 결과를 보면 방송통신 심의위원회에서 유해게시물로 판단되어 시정 요구 받은 64.446건 중 59,422건이 SNS에서 발생한 것으로, 전체 시정 요구들 중에서 92.2% 비율을 차지하고 있다. 이와 같은 조사 결과는 웹페이지 또는 인터넷 카페, 블로그 등을 이용하여 공유되어지는 유해정보의 수보다 SNS를 통해 공유되어지는 정보의 수가 훨씬 많으며, 실제 SNS상에서의 유해 또는 불법 정보의 유통문제의 심각함을 보이는 결과라고 할 수 있다.

또한 '불법음란물 판치는 구글-페북-인스타그램..대책은?' [2] 의 내용을 보면 외국계 포털 및 SNS 서비스에서 유통되는 유해정보를 차단하기 위해 해당 게시물의 유해함을 증명하는 증거자료를 수집하고, 해당 외국 기업에 제출한 뒤 게시물 시정을 요구하는 방법을 사용 하고 있음을 알 수 있다.

본 논문에서 제안하는 유해 해시태그 비율 기반의 유해 정보 판단 및 수집 시스템은 상기 서술한 SNS를 통한 불법 및 유해 정보 유통의 문제점을 해결하기 위해 다음과 같은 방법을 통해 유해 정보 판단을 수행한다.

1) SNS 게시물을 실시간으로 수집한다.
2) 수집된 게시물의 해시태그 용어 추출한다.
3) 추출된 해시태그 용어들과 구축된 유해 단어 DB를 비교하여 해시태그 용어의 유해 여부를 확인 한다.
4) 유해 용어로 확인된 해시태그의 비율을 계산 한다.
5) 수집 된 게시물에서의 유해 용어를 사용한 해시태그 비율이 과반수인 경우, 해당 게시물이 유해 정보를 포함하고 있다고 판단한다.

또한 논문에서는 제안하는 방법의 성능을 측정하기 위하여 실제 Twitter 데이터를 바탕으로 실험을 진행하였으며, 실험 결과에 대한 분석 방법으로 혼동행렬 사용한 분석을 실시하였다.

2. 관련 연구

본 장에서는 기존의 유해정보 판별 기술과 관련된 기존 연구들을 제시하고, 해당 연구의 한계와 본 논문에서 제안하는 방법의 차이점을 다음과 같이 서술한다.

김영수 등에 의해 연구된 '유해 텍스트 판별 기술' [3]에서는 전체 텍스트 기반의 분석 방법을 제시하였다. 상기 논문에서 제안하는 방법은 유해 텍스트 판별기와 학습 기반 유해 텍스트 판별기를 같이 사용하는 것을 특징으로 한다. 이때 유해 텍스트 판별기를 사용하여, 대규모 문서에서 유해정보를 판단하기 위해서는 문서 분류 모델

중 하나인 기계 학습 모델 먼저 생성되어야 한다는 점과 문서내의 외부 웹문서로 이동하는 링크가 있을시 접속이 불가피하고, 웹 문서내의 유해정보를 판별할 시, 특수문자 및 불용어를 찾아 제거해야하는 전처리과정이 필요하다는 문제점이 있다.

이러한 문제를 해결하기 위해 'SNS 기반 유해사이트 판단 및 수집 시스템' [4]에서는 SNS 게시물을 수집하고, 수집된 게시물에서 추출한 웹 문서의 유해도 여부를 판단할 시, 특수문자, 불용어를 제거하는 전처리 과정을 실시하지 않고 유해단어의 등장 빈도수를 통해 유해 여부를 판단하는 방법을 사용하였다. 하지만 유해 여부를 판단하려는 자료의 크기가 커질수록 수행시간이 크게 늘어나는 문제점과, SNS 게시물의 유해여부를 판단하는 기준인 유해단어의 등장 빈도의 기준치에 대한 산출 방법에 대하여 모호성을 지니고 있다.

본 논문에서 제안하는 유해 해시태그 비율 기반의 유해 정보 판단 및 수집 시스템은 전체 텍스트 또는 웹 문서를 바탕으로 유해 여부를 판단하는 것이 아닌, 사용된 해시태그 용어를 바탕으로 유해 여부를 판단하며, 웹 문서의 유해 여부를 판단하기 위해 링크 접속을 통한 해당 웹 문서 내용을 추출하는 과정과 특수문자 및 불용어를 제거하는 불필요한 전처리 과정을 제거하였다.

또한 텍스트 기반의 유해 여부 판단기법에서 실시하는 전체 텍스트에 대한 비교분석 방법과는 달리 해시태그를 대상으로 유해 여부를 판단하기 때문에 자료의 크기에 따른 수행시간의 증가율이 상대적으로 낮으며, 수집된 SNS 게시물의 유해 여부를 판단하기 위한 기계학습 모델 형성과 같은 사전 작업 과정이 존재하지 않는다.

그리고 SNS 게시물에 대하여 유해 여부를 판단하기 위해 유해 단어의 출현 비율을 사용함으로써, 앞서 언급된 유해 여부를 판단하는 기준의 모호성 또한 해결하였다.

3. 제안 기법

3.1 유해 여부 판단 기준

수집된 SNS 게시물의 유해 여부 판단 기준으로, 작성된 해시태그에서의 유해 용어 출현 비율을 사용한다. 전체 해시태그 중 유해 용어가 사용된 해시태그의 수가 과반수인 경우 해당 게시물을 유해하다고 판단한다. 이와 같은 방법을 사용하여 유해 여부를 판단하는 시스템의 흐름은 그림 1 과 같다.

3.2 유해 단어 데이터베이스 구축 방법

실시간으로 수집된 SNS 게시물의 해시태그들의 유해 여부를 판단하기 위해서는 유해 단어 데이터베이스의 구축이 매우 중요하며 해당 데이터베이스의 구축 방법은 다음과 같다.

1) 실제 유해하다고 판단된 SNS 유해 게시물들을 수집한다.
2) 수집된 유해 게시물들로부터 사용된 해시태그 용어들을 추출하고 이를 유해 단어로 분류한다.
3) 추출된 유해 단어들의 등장 빈도를 계산하고, 가장 높은 등장빈도를 기준으로 내림차순으로 단어를 정렬한다.
4) 가장 높은 등장 빈도를 가지는 단어를 기준으로, 약 1000개의 유해 단어들을 유해 단어 데이터베이스에 저장하고 이를 유해성 여부 판단에 활용한다.

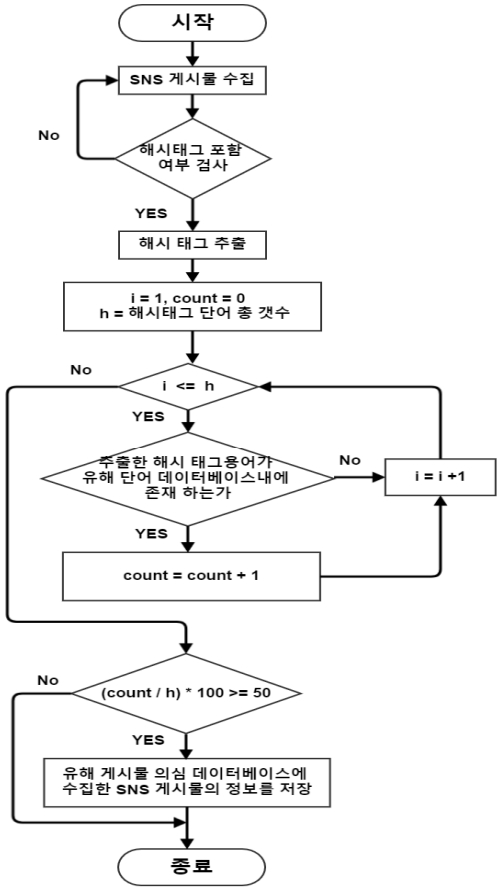


그림 1 제안하는 시스템의 흐름도

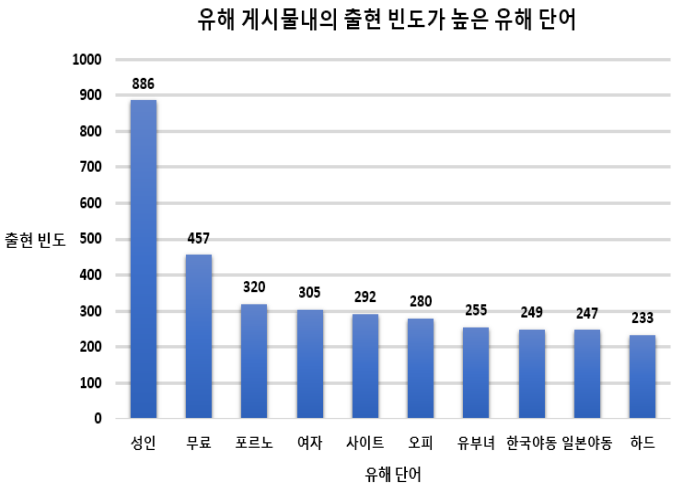


그림 2 등장 빈도에 따른 상위 10개의 단어

그림2 은 위와 같은 유해 단어 데이터베이스 구축방법에 의하여 구축된 유해 단어 데이터베이스로부터 가장 높은 출현 빈도를 보이는 상위 10개의 유해 단어와 각 유해 단어의 출현 빈도수를 막대그래프로 도식화한 그림이다.

4. 실험 조건 및 분석 방법

본 논문에서 제안하는 방법을 바탕으로 실제 실험에 사용된 프로그램은 Java로 개발되었으며, 실험 대상으로는 '소셜 미디어를 통한 불법·유해정보 유통 실태 및 대응 현황: 국내사례' [1]에서 가장 많은 유해 정보 게시물 시정요구를 받은 Twitter를 대상 SNS로 특정하여 실험을 진행하였다. 또한 구축된 유해 단어 데이터베이스와 비교를 수행할 실시간 SNS 게시물은 검색어로써 '야동'을 사용하였으며, 평가에 사용된 게시물의 수는 100개로 제한하였다.

실험은 정확도와 수행 시간을 평가하기 위해 총 두 가지의 실험을 차례로 진행하였다. 첫 번째 실험은 본 논문에서 제안하는 방법과 기존에 연구된 텍스트 기반의 방법을 사용하여 진행하였으며, 각각의 방법에 따른 결과를 혼동행렬을 통해 비교분석하였다. 두 번째 실험에서는 시험에 사용되는 게시물의 수를 100, 200, 300, 400, 500개로 증가시킨 경우의 프로그램 수행 속도 변화를 측정하였다.

5. 실험 결과

표 1 각 실험에 대한 혼동 행렬 분석 결과

	Accuracy	Precision	Recall	F_1
제안 방법	0.60	0.85	0.62	0.31
텍스트 기반 방법	0.66	0.82	0.76	0.08

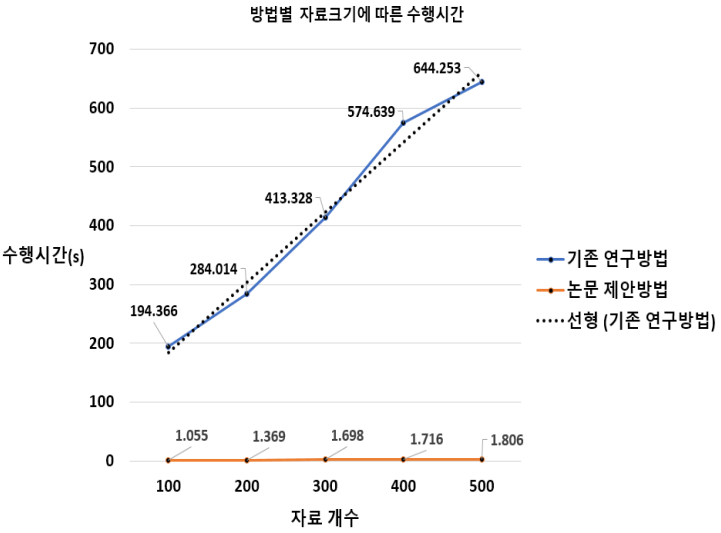
- Accuracy : 전체예측에서의 옳은 예측의 비율
- Precision : 예측에서의 옳은 예측의 비율
- Recall : 옳은 예측에서의 예측된 비율
- F1 : Precision과 Recall의 조화평균

본 실험에서 사용된 100개의 표본 SNS 게시물 중 실제로 유해한 게시물의 수는 83개였으며, 유해하지 않은 게시물의 수는 17개였다. 이와 같은 표본 데이터를 본 논문에서 제안된 방법을 통해 유해성을 판단한 결과, 61개의 게시물은 유해한 게시물로써 판단되었으며 39개의 게시물은 유해하지 않는 게시물로 판단되었다. 기존 연구인 텍스트 기반의 방법은 77개를 유해, 23개를 유해하지 않는 게시물로 판단하였다.

표1 은 각 방법에 대한 실험 결과를 혼동 행렬을 통해 분석한 결과이다. 본 논문에서 제안한 방법은 기존에 연구된 텍스트 기반방법에 비해 전체 예측에서의 옳은 비율과 실제로 유해한 게시물에 대하여 유해하다고 예측된 비율이 낮으나, 유해하다고 판단된 예측의 결과에 비율이 기존에 연구된 방법보다 높게 나왔다. 또한 방법의 성능을 평가하는 척도인 F_1측정의 결과 값이 0.31 으로써 기존 연구 방법의 0.08과 비교하여 향상된 성능을 보임을 알 수 있었다.

그림3 은 기존에 연구된 방법들과 본 논문에서 제안하는 방법을 사용하여 자료의 크기를 선형적으로 증가시킨 경우의 수행 시간 변화를 비교한 그래프이다. 기존 연구

그림 3 자료크기에 따른 수행시간 비교



방법은 자료의 크기에 따라 그 수행시간이 크게 증가하는 모습을 보이나, 제안하는 방법의 수행 시간은 자료 크기에 따른 수행시간의 증가폭이 매우 낮은 모습을 보임을 알 수 있었다.

6. 결론

본 논문에서는 기존 텍스트 기반의 유해성 판단 기법을 사용하였을 시, 자료의 크기에 따른 수행 시간의 증가와 유해 여부 판단에 앞서 모델 형성 및 전처리 과정 문제점을 해결하기 위해 해시 태그의 유해 용어 사용 비율을 기반으로 유해성 여부를 판단하였다. 제안된 방법은 기존에 연구된 방법들과 비교하여 자료의 크기에 따른 수행시간이 증가폭이 매우 낮아 보다 빠른 분석이 가능함을 보였다. 하지만 예측의 정확도의 측면에서는 기존 연구보다 낮은 수치를 보이고 있어 추후 추가적인 연구를 통해 예측의 정확도를 향상시키고자 한다.

ACKNOWLEDGEMENT

"본 논문은 교육부가 지원하고 충북대학교가 수행하는 지역선도대학육성사업의 지원을 받아서 수행되었습니다."

4. 참고 문헌

[1] 김경환 "소셜미디어를 통한 불법·유해정보 유통 실태 및 대응현황: 국내사례" 방송통신심의위원회 방송통신 심의동향, 2016

[2] 이경탁 "불법 음란물 판치는 구글-페북-인스타그램..대책은?" 아이티 투데이 http://www.kinews.net/news/articleView.html?idxno=67078

[3] 김영수, 남택용, 장종수 "유해 텍스트 판별 기술" 한국통신학회 학술대회 및 강연회, 345-349, 2005

[4] 장정현, 나스리디노프 아지즈 "SNS기반 유해사이트 판단 및 수집 시스템" 정보처리학회 춘계학술발표대회, 24권, 제 1호, 812-815, 2017