기계학습 기반 문장분석을 통한 악의적인 댓글 선별

Detecting malicious comments through machine learning based sentence analysis

저자
(Authors)

최봉근, 오민규, 설세형, 홍정우, 정은성
BongGeun Choi, Min Kyu Oh, Sehyung Seol, JungWoo Hong, Eun-Sung Jung

출처
(Source)

한국정보과학회 학술발표논문집 , 2019.12, 1463-1465(3 pages)

발행처
(Publisher)

한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09301967

APA Style

최봉근, 오민규, 설세형, 홍정우, 정은성 (2019). 기계학습 기반 문장분석을 통한 악의적인 댓글 선별. 한국정보과학회 학술발표논문집, 1463-1465

기계학습 기반 문장분석을 통한 악의적인 댓글 선별

최봉근 ○, 오민규, 설세형, 홍정우, 정은성
홍익대학교 소프트웨어융합학과
dhy00085@naver.com, omk9307@gmail.com, llbisull@naver.com, jungle0220@gmail.com,
ejung@hongik.ac.kr

Detecting malicious comments through machine learning based sentence analysis

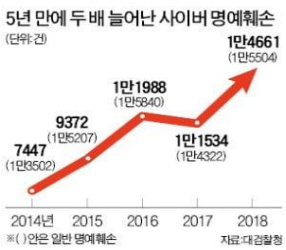
BongGeun Choi○ ,Min Kyu Oh , Sehyung Seol, JungWoo Hong, Eun-Sung Jung
Department of Software and Communications Engineering

요 약

2014년부터 2018년까지 사이버 명예훼손은 5년 동안 2배로 늘어났다. 이처럼 악의적인 댓글로 인한 피해는 점점 늘어나고 있는 추세이다. 하지만 많은 웹 사이트에서는 악의적인 댓글을 비속어를 키워드로 악의적인 댓글을 차단하거나 비속어에 해당하는 단어만 다른 언어 또는 xxx로 바꾸는 형식으로 차단하고 있다. 하지만 이런 차단방식은 비속어를 조금만 변경하면 이런 차단이 소용 없어지는 경우가 많다. 따라서 우리는 단어의 의미를 벡터화하는 워드임베딩 기술을 이용하여 기존에 차단하지 못했던 이런 변형된 비속어나 줄임 말 등을 판별하여 악의적인 댓글인지 의심적인 댓글인지 일반적인 댓글인지를 분류한다. 본 논문에서는 이런 변형된 비속어와 줄임 말이 사용된 악의적인 댓글을 어떻게 워드임베딩 기술을 사용해서 일반적인 댓글과 구별할 수 있었는지 프로그램 구성이 어떻게 되어 있는지에 대해 기술하였다

1. 서 론

인터넷이 활성화 되고 많은 커뮤니티 사이트들이 생기면서 많은 사람들의 접근이 쉬어졌다. 많은 사람들이 인터넷을 이용하다 보니 악의적인 댓글 즉 악플이 많은 문제를 야기해오고 있다.



왼쪽 표를 보면 대검찰청에서 발표한 2014년부터 2018년 까지 5년동안 사이버 명예 훼손이 2배로 늘어난 사실을 알 수 있다 이러한 악플은 너무 심해짐에 따라 악플 피해를 경험한 사람들은 우울증은 기본이고 심할 경우 자살까지 생각하는 경우가 있다. 얼마 전 설리가 자살을 한 것도 많은 사람들의 악플이 원인이라는 분석이 나오고 있으며 심지어 추모하는 글에도 악플이 달리는 등 심각한 문제를 야기하고 있다.

이러한 문제를 방지하기 위해 기존 사이트에서는 키워드를 이용한 악플 방지 시스템 또는 사용자들의 신고 시스템을 이용하고 있다 예를 들어 비속어를 쓴 댓글을 달려고 하면 자동으로 다른 문자로 바뀌어 올라가거나 또는 작성이 불가능하다. 다른 방법은 사용자들이 직접 보고 악플이라고 판단되는 문장을 신고를 해서 문장을 삭제 하는 시스템이다. 이러한 시스템은 단점이 명확한데 비속어가 작성이 안되니 중간에 숫자나 이상한 문자를 집어 넣어 시스템의 허점을 파고든다던가 다른 사용자들이 신고를 하지 않아 그대로 악플이 계속해서 남아 있는 경우가 많다. 따라서 이러한 문제점을 개선하는 방법을 생각하다가 단어의 의미를 벡터화 해서 비슷한 의미를 가진 단어를 비슷한 벡터로 표현하는 워드 임베딩 방식을 채택하여 오타가 있는 비속어와 줄임 말이 있어도 판별 가능하게 하였다.

우리는 워드 임베딩 모델을 이용해서 악의적인 댓글 선별 프로그램 구조를 설계 및 구현했다. 사용자는 선별하고 싶은 댓글에 해당하는 url을 입력하면 해당하는 댓글을 가지고 와 미리 학습시켜둔 워드임베딩 모델을 통해 선별한 후 전체 댓글, 악플, 의심, 일반 댓글로 나뉘어서 저장한다.

본 논문은 다음과 같은 구조로 작성이 되었다 2절에서는 워드 임베딩에 대한 배경 설명을 하고 3절에서는 악의적인 댓글 선별 프로그램의 전반적인 구현 구조 및 사용된 워드임베딩 모델의 설명과 학습시킨 결과 추가 기능을 자세히 설명한다 마지막으로 4절에서는 결론으로 논문을 마무리한다.

2. 배경 기술

일반적으로 악플을 차단 할 때는 미리 등록해둔 키워드로 차단하여 등록되지 않은 비속어나 변형된 비속어, 줄임 말을 차단하지 못한다. 워드 임베딩은 이러한 문제점들을 보완해줄 수 있는데 워드임베딩이란 단어를 벡터화하는 기술로 벡터에 단어의 의미를 담아 비슷한 의미의 단어들은 비슷한 벡터로 표현 된다. 예를 들어 '강아지'를 표현하는 벡터가 '멍멍이'를 표현하는 벡터와 얼마나 비슷한지, 또는 '연필'을 표현하는 벡터와는 얼마나 다른지를 벡터 간의 거리를 통해 알 수 있다.

이런 식으로 벡터에 의미를 담게 된다면 단어와 단어를 덧셈과 뺄셈도 가능하게 된다 예를 들면 "사랑 + 이별 = 추억" 이라는 결과를 받을 수 있다. 이렇게 단어를 벡터로 바꾸는 모델을 단어 임베딩 모델(word embedding model)이라고 부른다. 이 모델에서 대표적인 모델은 word2vec이며 이외에도 FastText, Glove 등이 있다.

3. 악의적인 댓글 선별 기능 구조 설계 및 구현

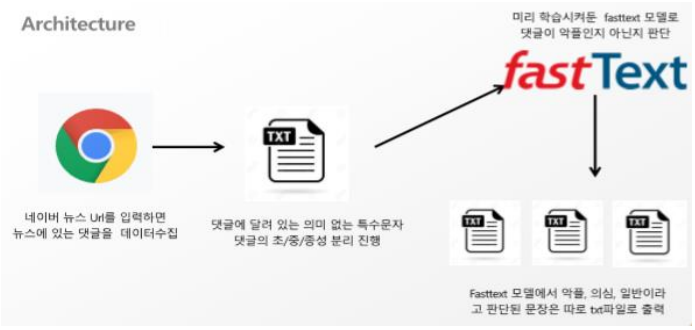


그림1 프로그램 구성도

이번에 개발한 프로그램은 변형된 비속어와 줄임말이 있어도 악플을 선별할수 있게 워드임베딩의 모델중 하나인 fastText를 이용하여 모델을 만들었다. 프로그램은 그림1 처럼 구성이 되어 있다. 하나의 기능을 예로 들면 해당 사용자가 악플을 선별하고 싶은 네이버 뉴스 url을 입력하면 뉴스에 있는 댓글을 수집한다 이렇게 수집한 댓글에 달려 있는 의미 없는 특수문자를 제거하고 댓글의 초중성을 분리하여 진행한다 이렇게 데이터 전처리가 된 파일을 미리 학습시켜둔 fastText모델로 댓글이 악플, 의심, 일반 이렇게 3가지로 선별하여 txt파일로 저장한다. 이런 과정이 끝나면 전체 댓글을 담고 있는 txt파일까지 해서 4가지의 txt파일이 생성이 된다. 2.배경기술에서 대표적인 모델이 word2vec라고 했는데 fastText를 사용한 이유는 밑에 자세히 기술했다.

3.1 word2vec

Word2vec의 핵심적인 아이디어는 '단어의 주변을 보면 그 단어를 안다' 이다. 보통 문장에 빈칸이 있을 때 단어의 주위를 보고 유추 할 수 있다. 빈칸에 들어갈 수 있는 단어들은 서로 비슷한 맥락을 갖는 단어들 즉 서로 비슷한 단어들이다. Word2vec는 predictive method라는 방식으로 비슷한 맥락에 비슷한 벡터를 주는데 predictive method란 맥락으로 단어를 예측하거나 단어로 맥락을 예측하는 문제를 지도학습처럼 해결 하는 것이다. word2vec에는 두 가지 예측 방식이 있는데 하나는 맥락으로 단어를 예측하는 CBOW모델이다 또 다른 하나는 단어로 맥락을 예측하는 skip-gram모델이 있다. CBOW모델과 skip-gram 모델은 서로 반대되는 개념이므로 CBOW모델만 살펴보면 맥락으로 타깃 단어를 예측하는 것이다 주변 단어란 타깃 단어의 직전 몇 단어 직후 몇 단어를 뜻하며 이것을 이용하여 벡터를 만든다 따라서 주변 단어 즉 맥락이 비슷하면 출력이 비슷해지면서 벡터간의 거리가 짧아지면서 서로 비슷한 의미를 가진 단어로 판별된다.

3.2 fastText

위에 설명대로 보자면 'word2vec로 충분히 프로그램을 구현할 수 있을 것 같은데 왜 fastText를 사용했는가'라는 질문이 충분히 나올 수 있다. 이 질문의 답변으로는 인터넷상의 댓글들이 문법을 철저히 지키지 않아 오탈 자와 문법이 파괴된 줄임 말 변형된 비속어 등이 난무하기 때문이다. Word2vec에서는 이러한 단어에 많이 취약하고 'out of vocabulary'라는 오류를 보여준다 따라서 이러한 문제를 해결하기 위해 찾은 방법이 fastText이다. fastText는 word2vec와 거의 비슷하지만 word2vec가 한 단어의 벡터 값을 학습한다면 fastText는 단어를 구성하는 subwords의 벡터 합으로 표현한다는 것이 다른 점이다. 예를 들어 '어디야'를 학습할 시 '어디야'가 아닌 [어디][디야]를 통해 학습을 한다 하지만 이렇게 되면 영어의 오탈 자는 잘 찾을 수 있을지 몰라도 한글 같은 경우는 힘이 들 수 있다 예를 들어 '어디야'를 '어딛야'라고 오타를 낼 경우 [어딛][딛야]라고 학습을 해서 '어디야'와 비슷한 단어로 인식되기 어려울 수 있다. 한국어의 오탈자는 보통 초/중/종성에서 한군데 정도가 틀리기 때문에 자모를 풀어서 학습을 한다면 이러한 문제를 해결할 수 있다. 따라서 학습 데이터를 초중종성 분리 전 처리가 필요해졌다.

3.3 악의적인 댓글 분류기준 및 학습시킨 데이터

모델에 데이터를 학습시키기 전에 저희 나름대로의 악의적인 댓글의 분류기준이 필요로 했다 따라서 저희는 나름대로 기준을 정했는데 공격성이 있으나 비속어가 없으면 '의심' 공격성이 있고 비속어도 있으면 '악플' 둘 다 없을 시 '일반'이라는 기준을 가지고 데이터를 준비했다 데이터는 네이버 뉴스 중 정치, 경제, 사회, 생활/문화, IT, 세계 부분에서 약 2017년부터 2019년 5월의 댓글을 가지고 와 직접 분류를 하였고 그렇게 해서 준비한 댓글이 1452483개의 댓글을 준비할 수 있었다.

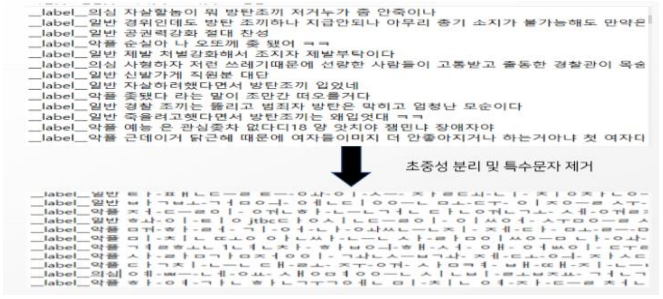


그림 2 데이터 전처리

그림 3 전처리후 정확도 변화

이렇게 준비한 데이터를 그림2처럼 초/중/종성 분리 및 특수문자를 제거했을 때와 제거하지 않았을 때를 비교해보니 정확도와 재현율 모두 96%되는 학습모델을 얻게 되었다.

3.4 추가기능 keyword 입력

그림 3 keyword 입력 기능 구성도

추가 기능인 keyword입력을 통한 악의적인 댓글 선별은 그림3처럼 구성되어있다. 이 프로그램은 댓글에 keyword가 있는 문장만을 검색을해 가지고 온다 그렇게 가지고 온 데이터를 기본기능과 같은방식으로 초/중/종성 분리와 특수문자 제거를 한뒤 fastText 모델에 넣어서 선별하는것이다. 왜 이런기능을 추가했는가에 대해서는 위에 설명한 기능들로 악플을 선별 할 수 있지만 악플 피해를 받은 사람들 중 너무 극심한 고통에 가해자를 고소하고 싶을 경우를 생각한것이다 모욕죄나 명예훼손에 대해 고소 하고 싶을 경우 주관적인 부분이 많지만 피해자의 이름이 들어가 있거나 피해자를 구분할수 있는 단어가 같이 작성될시 행당 죄에 고소가 가능하다 이때 keyword로 이름 또는 자신을 구분할수 있는 단어를 입력하면 거기에 관련된 문장만 가지고 온 뒤, 선별을 하기 때문에 가해자에 대한 처벌이 가능하다. 또한 댓글이 사람한테만 달리는 것이 아니다. 물건에 대해서도 댓글이 달린다. 이때 자사와 관련된 뉴스에는 당연히 자사의 물품에 관한 댓글이 있겠지만 경쟁사 물품에 대한 뉴스에도 자사의 물품에 관한 비판과 불만이 달릴수도 있다 이때 경쟁사 뉴스의 url을 입력하고 자사물품의 명을 입력하면 자사 물품에 대한 경쟁사와 비교나 불만을 볼 수가 있다.

3.5 악의적인 댓글 선별 프로그램 기능 구현

1)뉴스 댓글 크롤링

네이버 뉴스 url을 입력 받으면 해당하는 뉴스의 제목과 댓글을 가지고 온다 이때 댓글에 불필요한 특수문자를 제거하는 작업을 해서 한 문장씩 카운터 해 뉴스 제목과 댓글 수를 알려준다.

2) keyword입력 및 데이터 전처리

Keyword가 없을 시 전체댓글을 저장하지만 keyword가 있을 시에는 댓글에 keyword가 존재하는지 아닌지를 비교해서 keyword가 존재하는 문장만 '뉴스제목'.txt 파일에 저장한다. 이렇게 저장한 파일을 다시 한번 특수문자를 제거해주고 초/중/종성 분리를 해준다.

3) fastText모델에서 선별

데이터 전처리가 완료된 파일을 미리 학습시켜둔 fastText모델에 넣어 선별을 한다 이때 fastText에 문장을 입력하면 '악플', '의심', '일반', 이3가지 라벨 중 하나가 결과로 나오고 이 라벨에 따라 따로 저장해주면서 각자 몇 문장이 나왔는지 카운트 해준 다음 결과로 각각 몇 문장씩 있는지 보여주고 각자 저장을 한다.

4) 프로그램 시연 영상

키워드 없이 댓글 선별

키워드 있는 댓글 선별

경쟁사 뉴스에서 댓글 선별

4. 결 론

본 논문에서는 악의적인 댓글 선별을 위한 프로그램 설계 및 구현에 대해 기술하였다 선별 프로그램 구현을 위해 워드임베딩 모델 중 하나인 fastText를 이용하여 변형된 비속어나 줄임 말이 포함된 문장까지 선별 가능하게 만들었으며 keyword기능을 추가하여 피해자가 가해자를 고소할 수 있는 가능성을 열어두고 경쟁사 제품 뉴스에서 자사 제품에 대한 불만이나 비교할 점이 있는지 알아볼 수 있는 기능까지 구현하였다.

참 고 문 헌

[1]. Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, "Enriching word vectors with subword information", pp.1-12

[2]. Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov, "Bag of Tricks for Efficient Text Classification", pp.1-5

[3]. Word2Vec Tutorial - The Skip-Gram Model, http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

[4]. Word2Vec Resources, http://mccormickml.com/2016/04/27/word2vec-resources/

[5]. fasttext DOCS, https://fasttext.cc/