기계 학습을 이용한 악성 댓글 판별 시스템

신효정[0], 최소운, 이경호, 이공주
충남대학교
gomdoi48@hanmail.net, chlthdns009@naver.com, gyholee@gmail.com, kjoolee@cnu.ac.kr

Discrimination System for Abusive Comments using Machine Learning

Hyo-jeong Shin[0], So-Woon Choi, Kyung-ho Lee, Kong-Joo Lee
Chungnam National University

요 약

본 논문에서는 기계 학습(Machine Learning)을 이용하여 댓글의 악성 여부를 분류하는 시스템에 대해 설명한다. 댓글은 문장의 길이가 짧고 맞춤법이 잘 되어있지 않는 특성을 가지고 있다. 따라서 댓글 분석을 위해 형태소 분석 결과와 문자단위 Bi-gram, Tri-gram을 자질로 이용한다. 전처리 된 댓글에서 각 자질 추출 방법에 따라 자질을 추출한다. 추출된 자질을 이용하여 기계학습 알고리즘의 모델을 학습하고 댓글의 악성 여부 분류에 활용한다. 본 논문에서는 댓글의 악성 여부 판별을 위한 자질 추출방법을 제안하고 실험을 통해 이에 대한 효용성을 검증하였다.

주제어 : 악성 댓글, 서포트 벡터 머신, 백 오브 워즈, 워드 투 벡터, 군집화

1. 서론

인터넷의 성장과 스마트폰의 발전으로 사람들은 뉴스 댓글, SNS 등에 수많은 글을 올리고 자신의 의견을 남긴다. 다른 사람들이 남긴 평을 읽으며 공감하기도 하고 다른 사람을 비판하는 글을 쓰기도 하면서 서로의 의견을 공유한다. 다양한 의견을 쉽고 빠르게 공유할 수 있지만 그만큼 많은 사회적 문제점을 일으키고 있다. 악성 댓글이 이러한 사회적 문제 중 하나이다. 유명인을 향한 악성 댓글 뿐만 아니라, 개인 대 개인에 대한 악성 댓글의 분쟁은 온라인을 넘어 오프라인에서도 그 영향을 끼치고 있다. 본 논문에서는 이러한 문제점을 완화 할 수 있는 악성 댓글 판별 시스템의 구성을 제안한다.

2. 관련 연구

기존에도 서포트 벡터 머신(Support Vector Machine, SVM)을 이용해 악성 댓글을 판별하는 연구가 있었다[4]. 이 연구에서는 자질에 대한 출현빈도(Term Frequency)와 역 문헌빈도(Inverse Document Frequency)를 가중치로 하여 문서를 벡터화하였다. 수치화 된 데이터는 <자질:값> 형태로 표현되고 학습 데이터로 활용하였다. 학습 데이터를 통해 SVM모델을 학습시키고 그것을 기반으로 댓글을 분류하게 된다. 이 연구에서 댓글의 분류 정확도는 68.90%의 결과를 내었다.

본 논문에서는 백 오브 워즈(Bag of words, BoW), 워드 임베딩(Word embedding), 군집화(Clustering)방법을 이용하여 자질을 추출하였다. 다양한 자질 결정 방법을 통해 정확한 결과를 얻고자 하였다.

3. 본론

3.1 전체 시스템 구조

본 논문에서 제안하는 시스템은 기계학습 모델을 기반으로 악성 댓글과 비악성 댓글을 분류한다. 분류기 학습을 위해 먼저 댓글을 수집하고, 이를 전처리한다. 전처리된 댓글은 자질 추출과정을 통해 일정한 길이의 벡터로 표현된다. 이렇게 추출된 자질을 SVM 알고리즘[3]에 적용하여 기계학습 모델을 학습한다.

댓글 판별이 필요한 댓글이 입력되면, 앞선 모델 학습 단계와 같은 과정을 통해 자질을 추출한다. 추출된 자질을 학습된 모델에 적용하여 입력된 댓글의 악성 댓글 여부를 판별한다.

전체 시스템 구조는 그림 1과 같다.

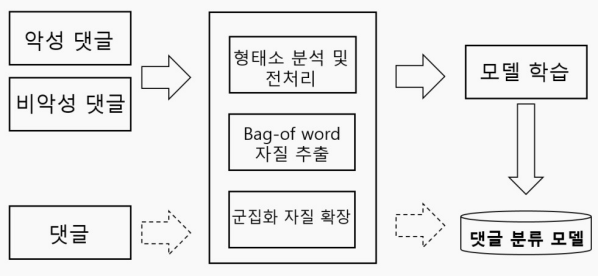

그림 1 . 전체 시스템 구조

그림에서 실선 화살표는 학습단계의 과정을 나타내고 점선 화살표는 분류 단계의 과정을 나타낸다.

3.2 데이터 수집

악성 댓글 판별을 위한 모델 학습 및 성능 평가를 위해 댓글을 수집하였다. 댓글은 1)인터넷 연예면 신문기사 댓글 2)온라인 커뮤니티 글의 댓글을 수집하였다. 1)의 댓글들은 신문기사의 특성상, 주로 시의성이 있는 다양한 여론의 댓글들이 달린다. 하지만 1)의 댓글의 경우, 심한 수준의 악성 댓글은 포털 및 신문사의 관리를 통해 제거가 된다. 2)의 댓글은 1)보다 덜 엄격한 수준에서 관리가 되는 인터넷 커뮤니티에서 수집하였다. 이를 통해 다양한 수준의 악성 댓글을 수집하였다.

3.3 자질(feature) 추출

3.3.1 전처리

온라인 상에서 사용하는 글들은 비 정형적이고 신조어, 사전에 등록되지 않은 말들이 많다. 이러한 단어들이 댓글의 악성 여부를 판단하는 데 중요한 부분을 차지한다. 그렇기 때문에 댓글에서 불필요한 요소를 삭제하여 자질을 추출하는 과정이 필요하다.

형태소 분석 결과 중, 자질로 사용한 형태소는 일반명사, 고유명사, 동사, 형용사, 일반부사이다. 또한 자음으로만 이루어진 단어 중 욕설로 많이 사용되는 단어를 자질에 포함하였다.

형태소 분석 결과, 형태소를 알 수 없는 것(Unknown, UNK)으로 분류된 단어들이 있다. 이는 형태소 분석기 사전에 등록되어 있지 않은 단어로 주로 신조어들이 많다. 인터넷 댓글에는 신조어가 많이 사용되므로 이러한 UNK가 중요한 자질이 될 수 있다. 그렇기 때문에 이를 적절히 처리하여 자질로 사용할 수 있는 방법이 필요하다. 본 논문에서는 이러한 UNK 단어를 문자 단위 bi-gram과 문자단위 tri-gram으로 쪼개어 자질로 사용한다. 문자단위 n-gram 자질의 예는 그림 2와 같다.

그림 2 . 문자단위 n-gram 예

```
이빠이가 -> [ '이빠' , '빠이' , '이가'  # bi-gram
             '이빠이' ,' 빠이가'  #tri-gram]
```

3.3.2 자질 추출

BoW 자질

전처리 과정을 거친 댓글로부터 자질을 추출한다. 첫 실험에서는 BoW를 통해 자질을 추출하였다. 수작업을 통해 분류된 데이터를 기준으로 악성과 비악성 2개의 클래스(class)가 생성된다. 각 단어별로 빈도수를 확인하고 빈도수가 2이상인 단어가 자질로 추출된다. 이러한 단어들 중, 악성과 비악성 클래스에서 공통으로 자주 나온 단어와 양 클래스에서의 발생 빈도가 비슷한 단어는 자질에서 제외하였다. 학습 모델을 생성할 때 자질의 값은 해당 자질의 존재여부이다.

K 평균 군집화 자질

발음은 다르지만 같을 뜻을 의미하는 단어들이 온라인에서 많이 사용된다. 예를 들어 '좋아'라는 단어를 '조아, 죠아, 쪼아, 져아' 등 다양한 방법으로 나타낼 수 있다. 같은 뜻을 의미하는 단어들이 다른 자질로 표현 되기 때문에 자질 공간을 낭비하게 된다. 낭비 되는 공간이 커질수록 자질 벡터의 중요한 요소들이 드물게 나타나게 된다. 그 결과 악성 댓글 분류 성능을 떨어뜨린다. 그래서 낭비되는 자질 공간을 축소하기 위해 워드 임베딩(Word Embedding)과 K평균 군집화(K-means Clustering)를 사용하여 자질을 추출하였다. K 평균 군집화와 워드 임베딩을 이용하여 자질을 추출하는 방법은 다음과 같다. 1) 먼저 워드 임베딩을 통해 각 단어를 벡터로 표현한다. 워드 임베딩의 특성에 따라, 비슷한 맥락에서 사용된 단어들은 비슷한 벡터공간상에서 위치하게 된다. 2) k 평균 군집화 알고리즘을 통해 벡터 공간상에서 유사한 단어들을 같은 군집으로 묶어준다. 3) 댓글에서 각 군집에 포함된 단어 포함 여부를 k 개의 벡터 공간상에 표현함으로써 자질을 나타낸다. 이를 통해 단어 자질을 확장할 수 있고 자질 공간의 낭비를 줄일 수 있다.

4. 실험결과 및 분석

4.1 실험 데이터

3.2절에서 수집한 댓글에 대하여 태깅을 수행하였다. 학습과 평가에 사용한 데이터 개수는 다음과 같다.

	악성	비악성	전체
학습	500개	500개	1000개
평가	528개	2587개	3115개

표 1 . 학습과 평가에 사용한 데이터 개수

BoW의 자질 개수는 2,727개이고, 빈도 수 2이상인 단어만 자질로 추출 하였다.

워드 임베딩을 하기 위해 뉴스 기사 댓글, 온라인 커뮤니티 사이트 댓글의 약 10만 개 문서를 활용하였고 워드 투 벡터(Word 2 vec)[5] 알고리즘을 사용하여 워드임베딩을 수행하였다.

K평균 군집화는 K를 100개, 500개, 1000개로 변화시키며 실험하였다. 학습데이터에 사용한 댓글의 평균 어절 개수는 4개 이다.

4.2 실험 결과

표 2는 각 자질 추출 방법에 대한 실험 결과이다. Bow는 3.3.2에의 **BoW 자질에서 설명한 자질을 사용한 실험이다. K 평균군집화는 3.3.2의 K평균 군집화 자질을 사용한 실험 결과이다.**

자질		정확도(%)
백오브워즈(BoW)		76.8218
K평균군집화 (K =)	100	**77.1429**
	500	71.1075
	1000	72.2823

표 2 . 각 자질 추출 방법의 실험결과

BoW에 비해 K평균 군집화의 정확도가 더 높을 것이라는 예상과 달리 K평균 군집화 중 K=100인 결과만 BoW보다 높은 정확도를 나타내었다. K=500 실험 결과가 제일 낮은 정확도를 나타내었고 K=100 실험 결과가 제일 높은 정확도를 나타내었다.

5. 결론

본 논문에서는 온라인 상의 댓글을 모아 악성 여부를 몇 가지 방법으로 판단하는 시스템을 제안하고 그 성능을 검증하였다.

제안한 시스템의 성능은 K평균 군집화를 이용한 실험이 가장 우수한 성능을 보였다. 자질이 크면 정확도가 올라갈 것이라는 예상과 달리 K값이 큰 경우 오히려 정확도가 떨어지는 결과를 얻을 수 있었다. 또 데이터 수집에서 수작업을 통한 분류가 정확하게 되어야 값도 정확하게 나오게 된다. 정형화 되어 있지 않은 데이터들로 자질을 추출하는 것 또한 정확도를 떨어뜨리는 원인이 되었다.

댓글에 반어법과 비유법을 구분 할 수 없어 아쉬움이 남는다. 따라서 향후 과제로 비정형화 데이터를 정형화에 가까운 데이터로 변환하여 정확도를 올리고 반어법과 비유법을 구분하여 판별할 수 있는 방법이 연구되어야 할 것이다.

참고문헌

[1] 민도식, 송무희 손기준, 이상조, “SVM 분류 알고리즘을 이용한 스팸 메일 필터링”, 한국정보과학회 2003년도 봄 학술발표논문집 제 30권 제1호(B), 2003.4, 552-554(3pages)

[2] 김묘실, 국민대학교, “SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현”, 2007

[3] Joachims, Thorsten, “Making large scale SVM learnig practical”, 1999

[4] 배민영, 차정원, “Topic Signature를 이용한 댓글 분류 시스템”, 한국정보통신학회논문지 제16권 제9호, 2043-2049 (7pages)

[5] Tomas Mikolov 외 3명, “efficient estimation of word representations in vector space”, 2013