Topic Signature를 이용한 댓글 분류 시스템
(Comments Classification System using Topic Signature)

배 민 영 † 차 정 원 ††
(Min-Young Bae) (Jeong-Won Cha)

요 약 본 논문에서는 토픽 시그너처(Topic Signature)를 이용하여 댓글을 분류하는 시스템에 대해서 설명한다. 토픽 시그너처는 자질을 선택하는 방법으로 문서요약이나 문서분류에서 사용하는 방법이다. 댓글은 문장의 길이가 짧고 띄어쓰기가 거의 없으며 특수문자들이 많은 특성을 가지고 있다. 따라서 우리는 댓글을 7개의 음절로 나누고 이를 다시 Tri-gram으로 나누어 분류의 기본단위로 본다. 이 Tri-gram을 토픽 시그너처를 이용한 학습 단위로 사용하고, 학습한 자질을 베이지안(Bayesian) 모델을 사용하여 분류한다.

다양한 방법의 모델과 비교·실험을 통하여 구현한 시스템의 성능이 기존의 방법보다 상승되었음을 실험결과를 통해 알 수 있었다.

키워드 : 악플분류, 기계학습, 토픽 시그너처, n-gram

Abstract In this work, we describe comments classification system using topic signature. Topic signature is widely used for selecting feature in document classification and summarization. Comments are short and have so many word spacing errors, special characters. We firstly convert comments into 7-gram. We consider the 7-gram as sentence. We convert the 7-gram into 3-gram. We consider the 3-gram as word. We select key feature using topic signature and classify new inputs by the Naive Bayesian method.

From the result of experiments, we can see that the proposed method is outstanding over the previous methods.

Key words : comment classification, machine learning, topic signature, n-gram

1. 서 론

인터넷은 전 세계적으로 다양한 연령이 사용하고 있는 대표적인 서비스로 세계의 다양한 사건 사고와 핫이슈들을 실시간으로 접할 수 있고, 사용자들에게는 즐거움과 감동을 전해 준다. 또한, 네티켓·네티즌·댓글 문화 등 다양한 신조어를 탄생시킴으로써 현대인의 문화 트렌드라 할 수 있다.

그러나 토론과 비판의 장이 되었던 인터넷 공간이 익명성을 악용한 범죄의 공간으로 변하고 있다. 경찰청이 제출한 '유형별 사이버범죄 발생 및 검거현황' 통계에 따르면 악성댓글에 의한 피해 건수는 2002년 3,155건에서 2006년 7,881건으로 약 2.6배 증가한 것으로 조사되었다. 정상적인 댓글의 트래픽보다 악성댓글의 트래픽이 매우 크다는 사실이 연구 결과 밝혀졌으며[1], 지난 2007년 개최 된 MIT Spam Conference의 주제 중 상당수가 스팸 관련 연구였다는 점에서 악성댓글의 관심과 심각성을 일깨워 준다[2].

현재 악성댓글을 방지하기 위해 연구된 방법들의 대부분은 품사 태거 혹은 명사추출기 등을 이용한 연구이다. 그러나 악성댓글의 경우 정상적인 댓글에 비해 길이가 매우 짧고, 대부분 띄어쓰기나 정확한 단어의 사용이 이루어지지 않는다는 특징이 있어 기존의 방식으로는 높은 성능을 기대하기 어렵다. 또한, 분류 과정에서 유행어나 변형어가 많아 주요어(Keyword)를 기반으로 분류 할 경우 오류 발생 확률이 높아져 분류의 정확성이

· 이 논문은 2008 한국컴퓨터종합학술대회에서 'Topic Signature를 이용한 댓글 분류 시스템'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 창원대학교 컴퓨터공학과
nikismy@changwon.ac.kr
†† 종신회원 : 창원대학교 컴퓨터공학과 교수
jcha@changwon.ac.kr

논문접수 : 2008년 8월 25일
심사완료 : 2008년 11월 7일

Copyright@2008 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

떨어지게 된다.

악성댓글의 경우, 비방을 위해 비속어가 눈에 띄게 많이 사용된다. 그러나 대부분의 사이트에서 비속어 사용을 금지하기 위해 1차적인 단어 정제(filtering) 방법을 이용하므로 띄어쓰기나 맞춤법에 어긋나는 비속어의 변형어가 증가하고 있다. 정제를 피하기 위해 '.' ',' '/' 등의 기호와 특수문자를 섞어 사용하거나 일어·중국어 등 외래어를 이용한 변형이 증가하여 이를 시스템이 모두 알기란 현실적으로 쉽지 않다. 표 1은 이러한 변형어의 다양한 사례이다.

표 1 단어 변형의 다양한 방법

	기존	변형	방법
예1	미친 새끼	미.친.. 새//ㄲ ㅣ	,./; 등의 기호와 특수 문자, 공백 이용
예2	씨 발 놈	ssi bal nom	한글의 발음 나는 대로 영어쓰기
예3	사람	ㅅㅏㄹㅂ	한글+일어

일반적인 댓글의 경우, 악성댓글과 달리 이러한 변형이나 맞춤법이 틀리는 등의 문제가 거의 나타나지 않는다. 따라서 일반댓글과 악성댓글의 특징을 기반으로 자질을 추출하고 악성댓글 여부를 판단하는 시스템을 구현하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 연구에 대한 조사를 하고 3장에서는 시스템에 사용 된 학습 데이타와 평가 데이타 및 악성댓글 문장 패턴 학습을 통해 구현된 시스템에 대한 소개를 한다. 4장에서는 다른 모델들과의 실험을 통해 본 시스템의 효율성을 확인하고 실험 결과를 바탕으로 시스템의 성능을 분석한다. 마지막으로 5장에서는 본 시스템에 대한 결론과 향후 과제를 제시한다.

2. 관련 연구

2.1 감정분류(Sentiment Classification)

악성댓글을 판단하는 것은 단어가 가지는 의미를 파악하는 것과 문맥상에서 그 단어가 어떤 의미로 사용되었는지(좋은지 혹은 나쁜지)를 판단해야 한다. 이처럼 문맥상에서 단어의 의미를 파악하고 작성자의 감정을 판단하는 분야가 감정 분류(Sentiment Classification)로 이와 관련한 연구가 국외에서는 활발히 진행되고 있다. 단순한 단어 일치에서 구, 문장, 문서로 범주를 넓혀가며 더 정확성 높은 연구가 이루어지고 있다.

2002년 Uni-gram과 Bi-gram, 품사(Parts Of Speech) 등 단어 중심 자질들을 이용한 실험[3]에서부터 2004년과 2005년 분류를 위한 자동 키워드 추출 및 현재 자질의 전·후 자질을 적용한 문장 단위의 분류[4,5]가 이루어졌다. 또한 2007년 제안된 시스템에서는 주변 문장의 확률을 바탕으로 문서의 감성을 분류한다[6]. 최근 감정 분류는 베이지안(Naive Bayes), 최대 엔트로피(Maximum Entropy), 지지 벡터 기계(Support Vector Machines) 등 기계학습을 이용한 연구가 활발하다.

2.2 국내외 악성댓글 분류

국내의 경우 악성댓글로 인한 사회적 문제가 대두되면서 이를 방지하기 위한 여러 가지 연구 및 방안들이 제시되고 있다. 기존의 악성댓글을 방지하기 위한 방법은 아래와 같다[7].

- HTML 태그 제한
- 댓글 등록을 위한 로그인
- IP 블랙리스트[8]
- 일정 시간 동안 동일 ID/IP 사용자의 덧글 등록 방지(throttling)
- Capcha를 이용한 Turing test[9]
- 오래된 글에 댓글 작성 제한
- 외부 링크를 내부 링크로 리다이렉트(re-direct)
- rel="nofollow" 태그를 사용[10]
- 기존 글과 동일한 언어로만 댓글 등록
- 언어 모델(Language Model) 이용[9]

대부분 제안된 방법은 사전(事前)에 악성댓글 등록을 차단하는 것으로 국내의 경우 역시 이러한 방식을 취하고 있다. 위더스정보는 악성댓글을 애초에 올릴 수 없도록 하는 시스템을 개발 2007년 10월 특허를 받았으며, 네이버의 경우 작성자의 작성글에 따라 점수를 부여하는 '클린점수제도'를 도입하였다. 또한, 지난 5년간 치열한 찬반 논란을 벌였던 '인터넷 실명제'가 2006년 10월 국회를 통과, 2007년 7월부터 주요 포털이나 언론 등 1일 방문자수 10만 명 이상의 사이트에 한해 제한적으로 시행 되고 있다. 2007년 10월 정보통신부의 보도 자료에 의하면 이 제도의 시행에 따라 악성댓글의 2.2%가 감소한 것으로 나타났다.

이와 반대로 등록된 댓글 중 악성댓글을 분류하기 위해 제안된 시스템으로는 카이 제곱 통계량(Chi-Square Statistic)을 기반으로 한 본문과 댓글의 동시출현 자질을 이용한 방법[11]과 지지 벡터 기계를 이용하여 문서의 자질 추출과 가중치 부여를 통한 방법[12]이 있다. 두 방법 모두 품사 태거 혹은 명사추출기를 이용, 특정 어절(자질)을 추출하는 방식으로 시스템을 구축하였다. 그러나 악성댓글의 경우 앞서 말한바와 같은 문제로 자질을 추출하는데 있어 오류 발생의 확률이 높아질 수 있다.

국외의 경우 2005년 www 컨퍼런스에서 스팸 댓글을 제거하기 위한 새로운 접근법의 논문이 Mishne, G.와

D. Carmel에 의해 발표 되었다[9]. 이 논문에서는 언어 모델을 이용하여 블로그의 본문과 댓글, 댓글이 링크된 페이지간의 유사도 비교를 통해 스팸 여부를 판단한다. 그러나 이 논문에서 제안된 방법론은 유사하거나 동일한 내용의 악성댓글이 연속적으로 등록되는 문제에 대해서는 처리하지 못한다.

3. 제안 댓글 분류 시스템

인터넷으로부터 수집된 댓글은 XML형식으로 만들며, 일반댓글의 경우 <title>과 <body> 부분, 악성댓글의 경우 사용자에 의해 수집된 특정 구간(<block>)의 내용을 학습하여 악성댓글 여부를 판별하는 시스템을 구축하였다.

시스템은 크게 학습 단계와 테스트 단계로 이루어지며, 시스템의 전반적인 구조는 그림 1과 같다.

학습과정에서는 댓글을 수집하여 N-gram으로 분리한다. N-gram을 이용하는 이유는 짧은 문장에서 너무 많은 주변 정보를 이용함에 따른 오류 발생의 문제를 최소화하기 위한 것이다. 본 논문에서는 7-gram을 사용한다. 그 이유는 악성댓글 구간 추출 과정에서 구간의 평균 문자수를 계산해 본 결과 7에 근접하는 값을 가졌기 때문이다. 이 7-gram을 다시 Tri-gram 단위로 나누어서 학습을 한다. 본 논문에서는 이 Tri-gram을 단어라 하고, 한 단어를 자질이라고 한다.

3.1 토픽 시그너처(Topic Signature)

신생 단어가 생겨나고 악성댓글의 문장 길이가 길지 않으므로 학습 데이타베이스를 구축하는데 있어 많은 어려움이 존재한다. 또한 일반댓글만을 학습할 경우 악성댓글보다 가능한 예들이 많으므로 현실에서 나타날 수 있는 모든 문장을 학습시키는 것은 불가능하다. 만약 악성 댓글만을 학습하여 자질을 추출할 경우 다음과 같은 몇 가지 문제점을 가질 수 있다.

- 신생 단어의 비 학습에 따른 처리 문제
- 악성댓글 내의 일반 글의 출현에 따른 높은 빈도수
- 악성댓글에서 사용되어지는 높은 빈도수의 자질 중 일반댓글에서 다른 의미로 사용 될 수 있는 경우

따라서 악성댓글과 일반댓글 모두를 학습한 후 자질을 추출하여 판별에 적용하는 토픽 시그너처를 이용한다. Chin-yew Lin[13]에 의해 제안된 Log-likelihood Ratio 기반의 토픽 시그너처는 단어 추출(Term Extraction) 방법을 사용한다.

학습을 위한 문서에서 자질을 추출하고, 각 자질이 어떤 문서집합(악성댓글/일반댓글)에서 얼마나 나타났는지에 대한 빈도수를 기록한다. 각 문서집합에서 나타난 총 자질의 수와 각 자질의 수를 이용하여 그 단어의 문서집합에서의 확률을 계산한다.

표 2의 테이블에서 토픽 시그너처는 식 (1)과 같다.

$$TS_s(t) = 2\times(v_{11}+v_{12}+v_{21}+v_{22}) \times\left(\frac{v_{11}}{(v_{11}+v_{21})\times(v_{11}+v_{12})}\right) \quad (1)$$

여기서 $TS_S(t)$는 단어 t가 SPAM에 속할 때 토픽 시그너처 값이다. 계산 된 각 단어에 대한 확률은 현재 카

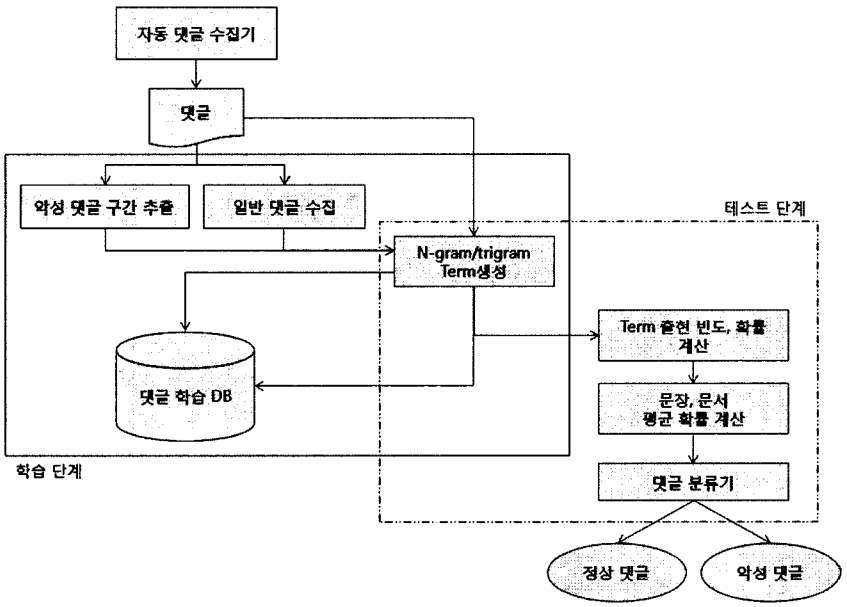

그림 1 패턴 학습 기반 댓글 분류 시스템 구조도

표 2 토픽 시그너처의 Contingency 테이블

	SPAM	NON-SPAM
t	v_{11}	v_{12}
~t	v_{21}	v_{22}

테고리에서 나온 횟수가 많고 다른 카테고리에서 나온 횟수가 적을수록 상위에 랭킹 된다. 최종적으로 순위화된 단어의 리스트를 이용해서 자주 나타나지 않은 단어(하위 순위)에 대해 평탄화(smoothing) 작업을 거쳐 학습 데이타베이스를 구축하게 된다.

3.2 학습 단계

학습 단계에서는 악성댓글의 <title>과 <body>부분 중 사람이 직접 선택 한 실제 악성댓글 구간만을 모아둔 <block>부분을 학습하게 된다. 댓글 수집 단계에서 중복되는 댓글을 배제하여 동일한 댓글이 반복 학습되는 것을 방지하였으며, 서로 다른 댓글에서 반복적으로 나타나는 악성댓글 구간은 학습 가능하도록 하였다.

또한, 악성댓글의 대부분이 띄어쓰기, 맞춤법을 고려하지 않고 비속어의 등록을 위해 단어 사이에 기호들(주로 . , /)과 공백을 사용하는 점을 감안하여 불필요한 기호와 공백을 제거한 후, 모든 단어를 한 문장에서의 문자 나열로 인식하였다. 표 3은 시스템에서 N-gram과 Tri-gram을 생성하는 과정이며, 여기서 생성된 Tri-gram은 식 (1)에 의해 확률이 계산되며 일정 확률 이상의 Tri-gram은 자질로 선택된다.

표 3 시스템의 'N-gram/Tri-gram 생성' 과정

입력	미소가 아름다운 사람은.. //마음//도 아름답다.		
문장	미소가아름다운사람은마음도아름답다		
N-gram (구간)	\t\t미소가아름다운\t	...	\t\t사람은마음도아\t
	\t\t소가아름다운사\t	...	\t\t람은마음도아름\t
	\t\t가아름다운사람\t	...	\t\t은마음도아름답\t
	\t\t아름다운사람은\t	...	\t\t마음도아름답다\t
Tri-gram	\t\t미 \t미소 미소가 소가아 아름다 름다운 다운\t		

3.3 평가 단계

학습된 데이타베이스를 이용하여 각 댓글의 카테고리를 결정한다. 각 댓글에서 자질을 추출하고 댓글에 나타난 자질에 대한 확률값을 학습 데이타베이스에서 찾는다. 식 (2)와 같이 베이지안(Naive Bayes) 모델을 사용하여 문서에 대한 카테고리의 확률을 계산한다. 이때, 확률을 문서에서 나타난 총 자질의 수만큼 나누어 문서의 길이에 따라 분류에 영향을 미치는 것을 방지한다.

식 (3)에서 $P(D|C)$는 카테고리에서 문서가 나타날 확률이고, C는 카테고리를 나타낸다. AC는 하나의 댓글에서 나타나는 총 자질의 수이다.

$$P(C|D) = \left(\frac{P(C)P(D|C)}{P(D)} \right) \quad (2)$$

```
<?xml version = '1.0' encoding='utf-8' ?>↓
<DATA>↓
<TITLE>↓
</TITLE>↓
<DATE>↓
2007.11.03↓
</DATE>↓
<ID>↓
hylee35↓
</ID>↓
<IP>↓
</IP>↓
<BODY>↓
과거2번이나밀어준유권자에,제님집안일로낙선,,있어버린10년,경제파탄,민생파탄안겨준원인제공자,,실
망만던져준맹한새까,무슨낯짜기로겨나오냐,?, 문노룰느낀다,겨나와,보수진영표만갈라놓는쳐주길넘,,,
나서야할중요한시기엔,쏘옥빠져숨엇다가,,나이72세나쳐먹은인간이,,좌파정권바꿀싯점에,,왜,겨나와서
찬물뿌리며,,역적노릇하려는가,,,대한민국에서,젤로싹아지없는놈/이인제 보다더싹아지없는쎄키웰려나,
,전에갖고있던대선자금수백억,,확실하게밝혀놓코,,한나라당에반납후,겨나가라,,,,,↓
</BODY>↓
↓
<BLOCK>↓
맹한새까 4↓
무슨낯짜기로겨나오냐 10↓
겨나와 3↓
쳐주길넘 4↓
젤로싹아지없는놈 8↓
싹아지없는쎄키 7↓
겨나가라 4↓
</BLOCK>↓
</DATA>←
```

그림 2 학습문서의 구조와 사람에 의해 선택 된 악성댓글 구간

$$P(D|C) = \frac{\prod_{i=1}^{AC} TS_{c,i}(t)}{AC} \quad (3)$$

판정이 어려운 구간(Gray Area)을 지정하지 않은 이유는 적은 가능성이라도 모든 경우를 판별해 내고자 하였기 때문이다.

4. 실험 및 토의

4.1 실험 환경 및 평가 방법

실험 데이타는 YAHOO Korea (http://kr.yahoo.com/)의 정치 뉴스 분야의 기사를 무작위 선택, 댓글을 수집하였으며, 악성댓글의 특정 구간은 사람이 직접 선택하였다. 시스템 평가를 위한 문서 집합은 표 4와 같다.

표 4 학습과 테스트에 사용 된 문서 집합

	NON-SPAM	SPAM		
		문서량	구간수	단어수
학습 데이타	1,000문서	1,000문서	1,734	11,350
평가 데이타	170문서	130문서	-	-
총 댓글	1,170문서 / 1,130문서			

성능 평가의 방법으로는 분류 시스템의 성능 평가를 위해 주로 사용되는 Precision, Recall, F_1-measure를 이용하였다. 구하는 공식은 식 (4)-(6)과 같다.

System \ Gold	SPAM	NON-SPAM
SPAM	A	B
NON-SPAM	C	D

$$P(Precision) : \frac{A}{A+B} \quad (4)$$

$$R(Recall) : \frac{A}{A+C} \quad (5)$$

$$F_1(F_1-measure) : \frac{2\times Recall\times Precision}{Recall+Precision} \quad (6)$$

4.2 실험 결과 및 토의

표 6은 토픽 시그너처, 카이 제곱 통계량에 대한 각각의 실험 결과이다. 실험 A는 문장을 Tri-gram으로만 나눈 방법이고, 실험 B는 문장을 N-gram으로 나누고 N-gram을 다시 Tri-gram으로 나눈 방법이다. N-gram으로 나누지 않고 Tri-gram만으로 나누는 방법은 표 5와 같다.

표 7은 실험 B의 방법에서 다음과 같이 실험 환경을 변경하여 실험한 것이다.

(1) 불용어를 제거하지 않은 경우

표 5 문장을 Tri-gram으로 생성하는 과정

입력	미소가 아름다운 사람은.. //마음//도 아름답다.
문장	\t\t미소가아름다운사람은마음도아름답다\t
Tri-gram	\t\t미 \t미소 미소가 소가아 아름다 … …… … 도아름 아름답 름답다 답다\t

표 6 토픽 시그너처, 카이스퀘어를 이용한 성능 비교 실험 결과

Model / 평가(%)	토픽 시그너처		카이 제곱 통계량	
	실험 A	실험 B	실험 A	실험 B
P	0.712500	0.772455	0.797203	0.728814
R	0.876923	0.992308	0.876923	0.992308
F_1	0.786207	**0.868687**	0.835165	0.840391

표 7 실험 B 방법을 이용한 각 모델의 성능 비교

Model / 평가(%)	(1)		(2)	
	TS	Ch-S	TS	Ch-S
P	0.753012	0.781513	0.699346	0.670968
R	0.961538	0.715385	1	1
F_1	**0.844595**	0.746988	0.823077	0.803089

표 8 Uni-gram과 Bi-gram을 이용한 성능 비교

Model / 평가(%)	Uni-gram		Bi-gram	
	TS	Ch-S	TS	Ch-S
P	0.738372	0.567442	0.72067	0.834646
R	0.976923	0.938462	0.992308	0.815385
F_1	**0.84106**	0.707246	0.834951	0.824903

(2) 악성댓글 전체를 학습한 경우

표 8은 문장을 N-gram으로 나눈 후 Uni-gram과 Bi-gram으로 나누어 실험한 것이다.

표 6과 표 7, 표 8을 통해 모든 모델이 74% 이상의 성능을 보임을 확인할 수 있다. Tri-gram을 이용할 경우 Uni-gram과 Bi-gram을 이용한 경우보다 더 좋은 성능을 보임을 표 6과 표 8을 통해 확인할 수 있다. 또한 토픽 시그너처를 이용한 실험에서 출현빈도에 따른 성능을 측정한 결과 출현빈도가 1일 때 더 좋은 성능을 보임을 실험을 통해 확인할 수 있었다.

기존의 연구에서 제안된 방법의 경우 대부분 명사나 특정 어절을 추출하는 방식을 사용하여 분류를 위해 선행되어야 하는 작업이 존재했다. 또한, 복잡하고 다양한 평가 방법에 따라 성능을 평가한다.

분류 속도의 문제도 발생했다. 영어의 경우 단어의 추출이 쉬운 반면, 한글의 경우 명사나 동사와 같은 특정 품사 추출을 위해서는 품사 태거 혹은 명사 추출기를 이용해야하는데 악성댓글의 경우 정형화되어 있지 않으므로 한계가 있다.

표 7의 비교모델의 추가 실험을 통해 확인할 수 있듯

이 위에서 언급한 다양한 악성댓글의 특징에 따른 문제점으로 높은 성능을 나타내지 못한다.

본 논문에서 제안 된 시스템은 선행 작업이 존재하지 않고, 단순히 패턴 매칭을 통해 분류하므로 악성댓글의 여러 특징에 따른 분류의 문제를 해결할 수 있었다. 또한 문장의 길이에 큰 영향을 받지 않으며, 변형어의 분류에도 높은 성능을 보였다.

5. 결론 및 향후 연구 과제

날로 증가하는 악성댓글은 이제 개인만의 문제를 넘어 사회 전반의 문제로 대두되었다. '익명', '온라인' 이라는 방식을 악용하여 타인을 비방하거나 불쾌감을 느끼게 하는 행동에 대한 잘못의 인식이나 반성이 절실히 요구되고 있으나, 해를 거듭할수록 악성댓글에 의한 피해 사례는 오히려 기하급수적으로 증가하고 있는 추세다.

이러한 악성댓글의 해결을 위해 다양한 연구가 진행되고 있으나 일반댓글에 비해 악성댓글은 그 길이가 매우 짧고, 비정형화 되어 있으며, 단어의 변형이 심하여 자동으로 악성댓글을 판별해 줄 수 있는 시스템의 개발이 쉽게 이루어지지 못하고 있는 실정이다.

본 논문에서는 악성댓글의 특징을 이용하여 단순한 패턴 매칭 방법을 이용한 방법이 악성댓글의 분류 성능을 개선할 수 있다는 것을 보였다. 정형화되지 않은 악성댓글의 다양한 패턴 학습을 통하여 기존 연구에 적용된 선행 작업들(품사부착, 특정 품사추출, 등)이 없이도 전체적인 시스템의 성능 향상이 가능함을 실험 결과로 보여준다.

또한 제안된 방법은 간단한 방법으로 이루어졌다. 다양한 형태의 악성댓글의 학습을 통하여 은유나 비유적인 표현으로 작성된 댓글도 분류해 낼 수 있었다.

본 논문에서는 댓글의 각 문장을 모두 N-gram으로 나눈 후 2차적으로 Tri-gram으로 나누어 Tri-gram의 출현 빈도와 확률을 계산하는 방식을 이용하였다. 그러나 대부분의 악성댓글의 경우 짧은 문장 길이에도 불구하고 특정 부분에 악성댓글임을 암시하는 단어나 문장이 존재했다.

만약 모든 문장을 N-gram으로 나누지 않고, 악성댓글의 특정 구간을 판별해 낼 수 있다면 더 빠른 속도로 악성댓글을 분류할 수 있는 시스템이 구현될 수 있을 것이라 생각된다.

참 고 문 헌

[1] comment and trackback spam statistics, http://akismet.com/stats/

[2] MIT Spam Conference 2007. http://www.spamconference.org/

[3] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP. pp.79-86. 2002.

[4] Soo-Min Kim and Eduard Hovy. Automatic Detection of Opinion Bearing Words and Sentences. IJCNLP. pp.61-66. 2005.

[5] Soo-Min Kim and Eduard Hovy. Determining the Sentiment of Opinions. COLING. pp.1367-1373. 2004.

[6] Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells and Jeff Reynar. Structured Models for Fine-to-Coarse Sentiment Analysis. EMNLP - CoNLL. pp.432-439. 2007.

[7] Spam in blogs, Wikipedia. http://en.wikipedia.org/wiki/Spam_in_blogs

[8] Movable Type Black Filter, with content filtering http://www.jayallen.org/projects/mt-blacklist/

[9] Mishne G., D. Carmel. Blocking Blog Spam with Language Model Disagreement. 1st International Workshop on Adversarial Information Retrieval on the Web. pp.1-6. 2005.

[10] Preventing comment spam using "nofollow" tag (2005). http://googleblog.blogspot.com/2005/01/preventing-comment-spam.html

[17] 전희원, 임해창. 본문과 덧글의 동시출현 자질을 이용한 역 카이 제곱 기반 블로그 덧글 스팸 필터 시스템. 한글 및 한국어 정보처리 학술대회 19th. pp.122-127. 2007.

[12] 김묘실, 강승식. SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현. 한글 및 한국어 정보처리 학술대회 18th. pp.285-289. 2006.

[13] Chin-Yew Lin and Eduard Hovy. The Automated Acquisition of Topic Signatures for Text Summarization. COLING 18th. pp.495-500. 2000.

배 민 영

2008년 창원대학교 컴퓨터공학과 졸업 (학사). 2008년 창원대학교 컴퓨터공학과 석사 재학 중. 관심분야는 자연어처리, 정보검색, 감정분류

차 정 원

2002년 포항공대 박사. 2002년~2003년 USC/ISI 박사후과정. 2003년~2004년 이화여대 전임강사. 2004년~현재 창원대학교 컴퓨터공학과 조교수. 연구분야는 자연어처리, 정보검색, 기계학습, 인공지능