Topic signature와 n-gram을 이용한 댓글 분류 시스템

배민영○ 차정원
창원대학교 컴퓨터공학과
nikismy@changwon.ac.kr, jcha@changwon.ac.kr

Comments Classification System using Topic Signature and n-gram

Min-Young Bae○ Jeong-won Cha
Dept. of Computer Engineering Changwon National University

요 약

본 논문에서는 토픽 시그너처(Topic Signature)와 n-gram을 이용한 댓글 분류 시스템을 개발한다. 토픽 시그너처는 문서요약이나 문서분류에서 자질 선택을 위한 방법으로 많이 사용되어지며, n-gram은 모든 언어에 적용 가능한 장점이 있다. 악성댓글은 대체로 문장 길이가 짧고 유행어나 변형어의 출현 빈도가 높으며 비정형화된 특징이 있다. 따라서 우리는 댓글을 n-gram으로 나누어 자질로 선택한다. 분류를 위해 베이지안(Bayesian)모델을 사용하였다. 본 논문에서는 한글과 영어 댓글에 대한 판별 실험을 통하여 구현한 시스템이 복잡한 전처리 과정이 필요한 기존에 제안된 방법들보다 더 나은 성능을 보이며, 언어에 관계없이 적용 가능하다는 것을 실험 결과를 통해 확인할 수 있었다.

1. 서 론

현대인의 문화 트렌드라 할 수 있는 인터넷으로 정보를 얻거나 서로의 의견을 주고받을 수 있는 기회가 확대되었다. 특히 블로그, 개인홈피 등의 활성화에 따라 누구나 자유로이 자신의 의견을 전달할 수 있는 댓글문화가 일반화되었다.

그러나 이러한 장점 이면에 이를 악용한 악성댓글 역시 사회적 문제로 대두되고 있다. 최근 잇단 연예인 자살의 주된 원인 중 하나로 악성댓글을 꼽고 있으며, '사이버모독죄'의 신설을 두고 찬반 토론 벌어지기도 했다. 한 조사결과에 따르면 'Akisme'라는 댓글의 악성여부를 알려주는 프로그램을 실행한 이후부터 현재까지의 댓글 성향을 분석한 결과 악성댓글이 비악성댓글에 비해 약 7배 이상 많은 것으로 나타났으며[1], 악플을 반대하는 '선플 달기 운동 본부'[2]가 출범하였다. 이처럼 악성댓글에 대한 관심과 연구가 꾸준히 증가하고 있다.

악성댓글을 차단하기 위해 IP 블랙 리스트를 유지[3]하거나 언어 모델(Language Model)을 이용[4]하는 등 다양한 방법이 제시되었으며[5], 주요 포털이나 언론 등 1일 방문자수 10만 명 이상의 사이트에 한하여 2007년 7월부터 '제한적 실명제'가 시행되고 있다.

악성댓글에 대한 다양한 연구 방법 중에서 대부분은 문서에서 주요어(Keyword)를 추출하여 자질로 사용한다. 그러나 악성댓글은 그 길이가 매우 짧고 무분별하게 사용된 단어가 많아 주요어를 이용할 경우 댓글의 악성여부를 판별하는데 한계가 있다. 또한 악성댓글 전체가 악성으로 판별되기 보다는 특정 구간이 악성이거나 비속어를 변형하여 사용하는 경우가 높다. 반면 비악성댓글은 문장 길이가 대체로 길며 잘못된 단어의 사용이 거의 나타나지 않는다. 그러나 비악성댓글만을 학습하기에는 그 형식이 너무 다양하므로 이 또한 학습되지 못한 많은 양의 단어 처리 때문에 문제가 발생할 수 있다.

따라서 비악성댓글과 악성댓글의 동시 학습을 통해 자질을 추출하고 댓글의 악성여부를 판별하는 시스템을 구현하고자 한다. 악성댓글의 경우 동일 단어의 출현빈도가 매우 낮으므로 낮은 출현빈도로도 문서 분류가 가능한 토픽 시그너처를 사용한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 연구에 대한 조사를 하고 3장에서는 댓글의 악성여부를 판별하는 시스템 구조와 자질 선택방법에 대한 내용을 설명한다. 4장에서는 다양한 방법의 실험을 통해 시스템의 성능을 분석한다. 마지막으로 5장에서는 결론 및 향후 연구과제에 대해 기술한다.

2. 관련 연구

댓글의 악성여부를 판별기 위해서는 단어가 가지는 문맥상의 의미를 판단해야 한다. 이와 같은 분야를 감정분류(Sentiment Classification)라 한다. 또한 같은 맥락에서 상위 연구 분야로는 문서 범주화(Document Categorization)가 있다.

감정 분류나 문서 범주화의 경우 국외에서 더 활발한

연구가 진행되고 있으며, 정확성 향상을 위하여 단순 단어 일치에서 구, 문장, 문서로 범주를 확대한 연구가 이루어지고 있다.

2002년에는 단어 중심 자질을 이용한 연구가 이루어졌으며[6], 이후 자질의 전후 단어를 이용하거나 주변 문장의 확률을 바탕으로 감성을 분류하는 연구가 이루어졌다[7, 8, 9]. 2005년 www 컨퍼런스에서는 악성댓글 제거를 위해 언어모델을 이용한 새로운 접근법을 Mishne,G와 D.Carmel이 제시하였다[4].

최근에는 베이지안(Naive Bayes), 최대 엔트로피(Maximum Entropy), 지지 벡터 기계(Support Vector Machines)[10, 11] 등을 이용한 연구가 활발하다. 이 방법들에서는 특정 영역별로 다르게 사용되는 단어 의미를 인지하고 그것을 분류한다[12, 13, 14, 15].

악성댓글 방지를 위해 국내에서도 다양한 연구와 시스템이 개발되고 있다. 2007년 10월 애초에 악성댓글을 올릴 수 없도록 하는 시스템을 위더스정보에서 개발하여 특허를 받았으며, 네이버는 '클린점수제도'를 도입하여 작성글의 악성여부에 따라 사용자에게 점수를 부여한다. 포털 사이트의 경우 인터넷실명제의 도입에 따라 악성댓글이 약 2.2%정도 감소했다는 정통부의 보고가 있었다.

등록된 댓글의 악성여부를 판별하기 위해 국내에서 제안된 방법으로는 다양한 자질의 추출하고 가중치를 부여한 후 지지 벡터 기계를 이용하는 방법[16]과 역 카이제곱 통계량(Inverse Chi-Square Statistic)을 기반으로 본문과 댓글의 동시출현 자질을 이용한 방법[17]이 있다. 두 방법 모두 품사태거 혹은 명사추출기를 이용하여 특정 자질을 추출, 이를 댓글의 악성여부 판별에 적용한다. 그러나 앞서 말한바와 같이 악성댓글은 비악성댓글과는 다른 몇 가지 특징을 가지므로 자질 추출에 있어 오류 발생 확률이 높아질 수 있다.

3. 악성댓글 분류 시스템

본 시스템의 이전 논문에서는 한글 댓글의 악성여부를 판별하기 위해 음절(character) 단위 n-gram을 이용한 실험이 이루어졌다[18]. 본 논문에서는 한글과 영어 댓글의 악성여부를 판별하기 위해 음절 단위와 단어(word) 단위 n-gram을 이용한 다양한 실험을 수행하였다.

본 시스템에 사용된 댓글 중 한글은 인터넷으로부터 직접 수집하여 XML형식으로 저장한다. 악성댓글의 경우 특정 구간(<block>)을 학습한 후 댓글의 악성여부를 판별하도록 시스템을 구축하였다. 영어의 경우 성능 비교를 위해 [17]에서 사용한 문서집합을 이용 하였다.

시스템은 크게 학습과 평가의 과정으로 이루어지며, 한글은 음절 단위의 n-gram(구간)으로 분리한 후 다시 음절 trigram으로 나누어 학습과 평가의 자질로 사용한다. 악성 댓글 구간 추출 과정에서 구간의 평균 음절수를 계산한 결과 7에 근접하는 값을 가진다는 것을 확인하였으며, 따라서 본 논문에서는 7-gram을 사용한다. 영어는 단어 단위의 n-gram으로 분리하여 학습과 평가의 자질로 사용한다.

3.1 토픽 시그너처(Topic Signature)

비악성댓글만을 학습할 경우 나타날 수 있는 자질의 종류가 악성댓글에 비해 많으므로 현실에서 나타날 수 있는 모든 문장 유형을 학습하는 것은 불가능하다. 또한 유행어가 즐비하고 문장 길이가 길지 않은 악성댓글의 경우 학습 데이터베이스를 구축하는데 많은 어려움이 존재한다. 악성댓글만을 학습할 경우 나타날 수 있는 몇 가지 문제점은 다음과 같다.

- 유행어의 비 학습에 따른 처리 문제
- 악성댓글 내의 일반 글의 출현에 다른 높은 빈도수
- 악성댓글에서 사용 되어지는 높은 빈도수의 자질 중 비악성댓글에서 다른 의미로 사용 될 수 있는 경우

따라서 본 시스템에서는 악성댓글과 비악성댓글 모두를 학습하여 판별을 위한 자질을 추출하는 토픽 시그너처를 이용한다. Chin-yew Lin[19]에 의해 제안된 Log-likelihood Ratio 기반의 토픽 시그너처는 단어 추출(Term Extraction) 방법을 사용한다. 짧은 문장 길이와 많은 유행어를 가지는 악성댓글의 경우 반복되는 단어의 사용이 적으므로 문서분류에 더 많이 사용되는 카이스퀘어(Chi-Square)보다 나은 성능을 기대해 볼 수 있다.

학습을 위해 문서에서 자질을 추출하고, 각 자질의 각 문서집합(악성댓글/비악성댓글)에서의 출현빈도를 기록한다. 각 문서집합에서 나타난 총 자질의 수와 각 자질의 수를 이용하여 그 자질의 문서집합 내 확률을 계산한다.

표 1. 토픽 시그너처의 Contingency 테이블

	악성댓글	비악성댓글
t	v_{11}	v_{12}
~t	v_{21}	v_{22}

[표 1]의 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 악성댓글에 속할 경우 토픽 시그너처 값이며, 계산된 각 단어에 대한 확률은 현재 문서집합에서의 출현빈도가 높고 반대 문서집합에서의 출현빈도가 낮을수록 상위에 위치한다. 최종적으로 순위화된 단어의 리스트를 이용해서 자주 나타나지 않은 단어(하위 순위)에 대해 평탄화(smoothing) 작업을 거친 후 학습 데이터베이스를 구축하게 된다.

3.2 학습 단계

본 논문에서는 한글과 영어 두 언어에 대한 댓글의 악성여부 판별하기 위해 두 언어에 대한 학습 방법을 약간 달리 하였다.

한글의 경우 악성댓글의 <title>과 <body> 부분 중 사람이 직접 선택 한 실제 악성댓글 구간만을 모아 둔 <block>부분을 학습 한다. 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 음절 나열로 인식하였다. [표 2]는 한글 시스템에서 음절 구간 7-gram과 음절 trigram을 생성하는 과정이며, 여기서 생성 된 음절 trigram은 식 (1)에 의해 확률이 계산되고 일정 확률 이상의 음절 trigram은 자질로 선택된다.

표 2. 한글의 음절(character) 'n-gram/trigram 생성' 과정

입력	미소가 아름다운 사람은.. //마음//도 아름답다.		
문장	미소가아름다운사람은마음도아름답다		
n-gram (구간)	₩t₩t미소가아름다운₩t ₩t₩t소가아름다운사₩t ₩t₩t가아름다운사람₩t ₩t₩t아름다운사람은₩t	… … … …	₩t₩t사람은마음도아₩t ₩t₩t람은마음도아름₩t ₩t₩t은마음도아름답₩t ₩t₩t마음도아름답다₩t
trigram (자질)	₩t₩t미 ₩t미소 미소가 소가아 아름다 름다운 다운₩t		

영어의 경우 한글과 마찬가지로 문자 n-gram으로 자질을 생성한 경우에 대한 실험을 수행 하였다. 또한 불필요한 기호들을 제거한 후, 문장을 공백 기준으로 분리하여 그 하나를 단어(word)라 지칭하며 n개의 단어를 연속하여 자질로 생성한 단어 n-gram에 대한 실험도 이루어 졌다.

3.3 평가 단계

학습된 데이터베이스를 이용하여 각 댓글의 문서집합을 결정한다. 각 댓글에서 자질을 추출하고 댓글에 나타난 자질에 대한 확률값을 학습 데이터베이스에서 찾는다. 식 (2)와 같이 베이지안(Naive Bayes) 모델을 사용하여 문서에 대한 카테고리별 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.

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

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

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

4. 실험 및 토의

4.1 실험 데이터 및 평가 방법

학습과 평가를 위해 사용된 데이터들은 한글과 영어 댓글로 한글 데이터의 경우 YAHOO Korea (http://kr.yahoo.com/)의 정치 뉴스 분야 기사를 다른 기간 동안 무작위로 선택하여 댓글을 수집하였으며, 악성댓글의 특정 구간은 사람이 직접 선택하였다. 영어 데이터의 경우 [17]에서 사용한 문서집합을 이용하였다. 시스템 평가를 위한 문서집합은 [표 3]과 같다.

표 3. 학습과 테스트에 사용 된 문서집합

	한글 데이터				영어 데이터	
	비악성 댓글	악성댓글			일반 댓글	악성 댓글
		문서량	구간수	단어수		
학습 데이타	1,200	1,200	2,047	13,553	10,000	19,586
평가 데이타	170	130	-	-	329	612
총댓글	1,370문서 / 1,330문서				10,329 / 20,198	

성능 평가 방법으로는 Precision, Recall, F_1-measure를 이용하였다. 구하는 공식은 식 (4)-(6) 와 같다.

System \ Gold	악성댓글	비악성댓글
악성댓글	A	B
비악성댓글	C	D

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

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

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

4.2 실험 결과 및 토의

[표 4]는 음절 단위 n-gram을 자질로 추출하고 토픽 시그너처와 카이 제곱 통계량의 값을 이용하여 댓글의 악성여부를 판별한 실험 결과이다. [표 4]의 실험은 문장을 구간(n-character)으로 나눈 후 다시 n-gram으로 잘라 이 n-gram을 자질로 선택한 경우이다. 한글의 경우 사용자에 의해 따로 수집한 악성댓글 구간(<block>)만을 학습한 후 음절 구간 7-gram, 음절 3-gram에 대한 실험 결과이며, 영어는 문자 구간 15-gram, 문자 7-gram에 대한 실험 결과이다.

표 4. 토픽 시그너처, 카이스퀘어를 이용한 성능 비교 실험 결과. 한글은 음절 3-gram, 영어는 문자 7-gram을 사용한 경우

평가(%) \ Model	Topic Signature		Chi-square	
	한글	영어	한글	영어
P	0.772455	0.981949	0.772152	0.604343
R	0.992308	0.420402	0.938462	0.774343
F_1	**0.868687**	0.588745	0.847222	0.678862

한글의 경우 [18]에서 실험한 결과에 따라 가장 높은 성능을 나타내는 음절 구간 7-gram, 음절 3-gram을 이용한 방법으로 실험을 수행하였다.

표에서 보면 한글의 경우는 띄어쓰기의 오류를 극복하기 위해 사용한 음절 n-gram이 적절하게 악성댓글을 구별하고 있다. 그러나 영어의 경우는 문자 n-gram으로는 의미 있는 결과를 내지 못하고 있다.

[표 5]와 [표 6]은 한글과 영어 댓글 문서에 대해 문장을 공백으로 나누어 단어(word) 단위 n-gram을 자질로 선택한 경우로 unigram, bigram, trigram을 각각 자질로 선택하여 실험한 결과이다.

[표 5]는 한글 단어 단위 n-gram에 대한 실험 결과로 학습 과정에서 악성댓글 구간을 학습한 경우(A)와 댓글 전체를 학습한 두 가지 경우(B)에 대한 실험 결과이다.

표 5. 한글 댓글에서 단어(word) unigram/bigram/trigram 자질을 이용한 실험결과

평가(%) \ Model			unigram	bigram	trigram
A	TS	P	0.773585	0.653846	0.666667
		R	0.630769	0.261538	0.215385
		F_1	**0.694915**	0.373626	0.325581
	CH-S	P	0.735294	0.631579	0.627451
		R	0.576923	0.276923	0.246154
		F_1	0.646552	0.385027	0.353591
B	TS	P	0.642384	0.542169	0.603175
		R	0.746154	0.346154	0.292308
		F_1	0.690391	0.422535	0.393782
	CH-S	P	0.531250	0.534653	0.586667
		R	0.784615	0.415385	0.338462
		F_1	0.633540	0.467532	0.429268

[표 6]의 왼쪽은 본 논문에서 제안하는 영어 단어 n-gram에 대한 실험 결과이며, 오른쪽은 동일한 영어 댓글 문서를 [17]에서 tf-idf와 '역카이제곱+동시 출현 자질'을 이용한 실험 결과이다.

표 6. 영어 댓글에서 단어(word) unigram/bigram/trigram 자질을 이용한 실험 결과

평가(%) \ Model		unigram	bigram	trigram	[17]
TS	P	0.918251	0.768448	0.877934	0.9566
	R	0.746522	0.933539	0.578053	0.6127
	F_1	0.823529	**0.842987**	0.697111	0.7470
CH-S	P	0.622093	0.610048	0.610048	
	R	0.826893	0.788253	0.788253	
	F_1	0.710020	0.687795	0.687795	

기존의 연구에서 제안되었던 방법은 대부분 특정 주요어를 추출하기 위해 선행 작업을 필요로 했다. 그러나 주요어 추출 과정에서 발생 가능한 오류가 악성여부 판별의 성능을 저하시킬 수 있다. 또한 대부분의 시스템이 복잡하고 다양한 분류방법을 복합적으로 사용하는 경우가 많았다. 본 논문에서는 n-gram을 이용하여 자질을 선택하므로 별도의 선행 작업을 요구하지 않는다.

각 실험을 통해 토픽 시그너처를 이용한 실험 결과가 카이스퀘어를 이용한 것보다 더 나은 성능을 보임을 확인할 수 있었다. 또한 한글과 영어 댓글에 상관없이 댓글의 악성여부를 판별하는데 84% 이상의 성능을 보임으로써 n-gram을 이용하는 방법이 특정 언어에만 적용되는 것이 아니라 모든 언어에 적용 가능함을 확인할 수 있었다. 이는 선행 작업에서 발생 가능한 오류가 없고 단순 패턴 매칭을 통한 분류가 이루어지므로 악성댓글의 여러 특징에 따른 분류 문제를 해결할 수 있다는 것을 나타낸다. 또한 문장 길이에 큰 영향을 받지 않으며 변형어나 유행어 등 새로운 패턴에 대한 처리도 가능하다는 것을 확인할 수 있었다.

한글의 경우 공백과 불필요한 기호를 제거한 후 2단계의 분리 작업을 거치는 음절 단위 n-gram 쪽이 더 좋은 성능을 보였으며, 영어의 경우 불필요한 기호를 제거한 단어 단위 n-gram이 더 높은 성능을 보였다.

이는 실험에 사용된 한글 문서의 경우, 반복적으로 올라오는 댓글의 학습을 차단하고 사용자에 의해 선택되어진 실제 악성댓글 구간만을 학습하므로 그 정확성이 더 높아졌기 때문으로 보여 진다. 또한 한글의 경우, 띄어쓰기가 없어도 문장을 이해하는데 큰 어려움이 없으므로 많은 인터넷 사용자들이 제한된 글을 쓸 수 있는 댓글에는 띄어쓰기를 하지 않는 특징이 있어 수집된 댓글의 대부분에도 띄어쓰기가 되어 있지 않는 경우가 많았기 때문인 듯하다.

영어 댓글의 경우, 단어 bigram의 경우 가장 높은 성능을 나타냈으며([표 6] 참조), 한글과 같이 공백을 제거하고 두 단계의 분리 과정을 거치는 방법의 경우 카이스퀘어를 이용한 경우 67.9%로 매우 낮은 판별 성능을 보였다([표 4] 참조). 이는 영어는 띄어쓰기가 없는 문자 나열만으로는 문장에 대한 이해가 어려우므로 대부분의 댓글에는 단어 사이의 띄어쓰기가 잘 이루어져 있기 때문인 듯하다. 또한 실험에 사용된 영어 댓글 문서의 특징상 악성댓글의 경우 연속되는 웹사이트 주소가 많아 특수기호와 공백을 제거한 단순문의 경우 문서 분류에 있어 사용될 자질로서 충분하지 못한 것으로 보여 진다.

5. 결론 및 향후 연구 과제

사회적, 법적 제제에도 불구하고 날로 증가하는 악성댓글은 이제 개인만의 문제를 넘어 사회 전반의 문제로 심각성이 증가하고 있다. 얼굴이 보이지 않는 공간에서 익명과 자유라는 이름하에 타인을 비방하고 불쾌감을 주는 행동에 대한 반성이 절실히 요구되고 있으나, 해를 거듭할수록 악성댓글로 인한 크고 작은 피해 사례는 오히려 전 세계적으로 급증하고 있는 추세다.

이러한 악성댓글로 인한 문제를 해결하기 위해 국내외적으로 많은 연구와 시스템이 개발되어지고 있으나 일반 댓글에 비해 그 길이가 매우 짧고 비정형화되어 있으며, 변형어나 유행어, 비유와 은유를 통해 악의적으로 댓글을 다는 경우가 많아 악성댓글을 판별하는 시스템을 개발하기란 쉽지 않은 실정이다. 사전에 악성댓글을 달 수 없도록 하기 위한 많은 방법이 제시되었으나 이 또한 큰 효과를 얻지 못하고 있다.

본 논문에서는 악성댓글의 특징을 이용하여, 단순 패턴매칭을 통해 댓글의 악성여부를 판별할 수 있다는 것을 보였다. 기존 연구에서 이루어진 선행 작업(품사부착, 특정 품사추출, 등)이 필요 없고, 간단한 방법을 이용하여 전체적인 시스템의 성능 향상이 가능함을 실험 결과로 확인할 수 있었다.

n-gram의 방식을 적용하여 특정 언어에 상관없이 모든 언어에 적용 가능하므로 범용성 또한 보여주었다. 이러한 n-gram의 방법은 다양한 형태의 악성댓글 학습을 보다 효율적으로 가능하도록 하였으며, 비유나 은유적으로 작성된 댓글에서도 비교적 잘 동작하였다.

본 논문에서는 음절 수준에서의 n-gram과 단어 수준에서의 n-gram을 각각 자질로 선택하여 자질의 출현 빈도와 확률을 계산하는 방식을 이용하였다. 동일한 단어가 자주 발생하지 않는 악성댓글의 특징상 카이스퀘어 대신 토픽 시그너처 값을 이용함으로써 더 나은 성능을 얻을 수 있었다.

한글의 경우, 두 단계의 분리 과정을 거쳐야하므로 악성댓글 중 실제 악성일 가능성이 높은 특정 구간을 판별해 낼 수 있는 전처리 과정이 있다면 더 빠른 속도로 댓글의 악성여부를 판별해 낼 수 있을 것이라 생각된다.

지지 벡터 기계와 같은 다른 분류 알고리즘을 적용하여 시스템의 성능을 평가하는 것은 추후 과제로 남겨두었다.

참 고 문 헌

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

[2] 선플 달기 운동 본부 http://www.sunfull.or.kr/index.php

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

[4] 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.

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

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

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

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

[9] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts. ACL. pp.271-278. 2004.

[10] Joachims, T. Text categorization with Support Vector Machines: Learning with Many Relevant Features. In Machine Learning. ECML-98. pp.137-142. 1998.

[11] P.D. Turney and M.L. Littman. Unsupervised Learning of Semantic Orientation from a Hundred-billion-word Corpus. No.ERB-1094 NRC #44929. 2002.

[12] Sara Owsley, Sanjay C. Sood, and Kristian J. Hammond. Domain Specific Affective Classification of Documents. AAAI. pp.181-183. 2006.

[13] Hong Qu, Andrea La Pietra, Sarah Pooh. Automated Blog Classification: Challenges and Pitfalls. AAAI. pp.184-186.. 2006.

[14] Michael Gamon. Sentiment Classification on Customer Feedback Data: noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings the 20th, International Conference in Computational Linguistics. pp.841-847. 2004.

[15] P.D. Turney and M.L. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM TOIS. Vol.21, No.4. pp.315-346. 2003.

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

[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.

[18] 배민영, 차정원. Topic Signature를 이용한 댓글 분류 시스템. 한국정보과학회 2008 종합학술대회 논문집 제35권 제1호(A), pp.81-82. 2008.

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