인스타그램에서 학습된 단어와 게시글의 메타 정보를 이용한 페이스북 텍스트 스팸 필터링 알고리즘

Facebook text spam filtering based on keywords learned from Instagram and meta-information of Facebook posts

저자
(Authors)

Junhong Kim, Deokseong Seo, Haedong Kim, Pilsung Kang

출처
(Source)

대한산업공학회 추계학술대회 논문집 , 2016.11, 1902-1927 (26 pages)

발행처
(Publisher)

대한산업공학회
Korean Institute Of Industrial Engineers

URL

http://www.dbpia.co.kr/Article/NODE07059996

APA Style

Junhong Kim, Deokseong Seo, Haedong Kim, Pilsung Kang (2016). 인스타그램에서 학습된 단어와 게시글의 메타 정보를 이용한 페이스북 텍스트 스팸 필터링 알고리즘. 대한산업공학회 추계학술대회 논문집, 1902-1927.

인스타그램에서 학습된 단어와 게시글의 메타 정보를 이용한 페이스북 텍스트 스팸 필터링 알고리즘

Facebook text spam filtering based on keywords learned from Instagram and meta-information of Facebook posts

Junhong Kim, Deokseong Seo, Haedong Kim, *Pilsung Kang

School of Industrial Management Engineering

Korea University

{junhongkim, heyhi16, haedong_kim, pilsung_kang}@korea.ac.kr

◆ Purpose of the research

1903

◆ Purpose of the research

Facebook 유저라면..

1904

◆ Purpose of the research

Facebook 유저라면..

**거의 모두 공감하여 불필요한
정보라고 판단되는 것이 있습니다.**

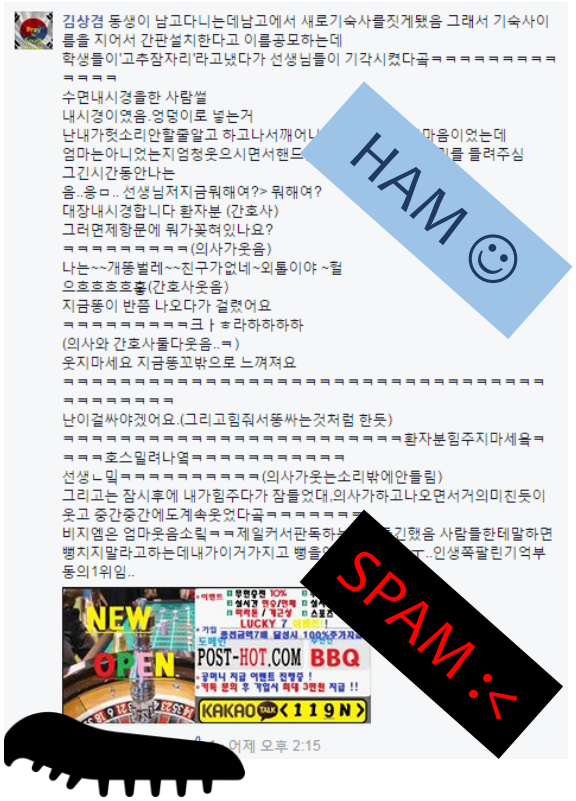
1905

◆ Purpose of the research

- 최근, 저비용으로 확산속도가 빠른 SNS기반의 스팸 광고가 활성화 되고 있음
- 페이스북의 유명 페이지의 포스트에 작성되는 댓글과 사용자가 올린 포스트중 스팸성 **댓글과 포스트**를 필터링 하고자 함
- 오른쪽 예시와 같이 text기반에서는 Ham 이지만 Image기반에서는 Spam으로 분류되는 고도화된 게시물이 생겨나고 있음

Case1. Text Spam & Image Spam
Case2. Text Spam

Case3. Text Ham, Image Spam



1906

〈연구 동기〉

Facebook 사용자의 입장에서 Spam의 정보가 무분별하게 노출되어 있으며, 한글기준 이를 차단하는 알고리듬이 구현되어 있지 않음

〈연구 방향〉

누구나 가용할 수 있는 공개된 데이터를 수집하여 Facebook의 Spam성 텍스트를 차단하는 연구방법론 개발

0. Research Framework

1 데이터 수집 및 전처리

원천 데이터

- 인스타그램 (Instagram)

데이터 특성	데이터 수	Hash Tag 개수
스팸 (Spam)	449,721	80 가지
정상 (Ham)	287,140	70 가지

- 페이스북 (Facebook)

데이터 특성	데이터 수	특이사항
페이스북 텍스트 데이터	1,795,067	23가지 페이지 2014.01.01. ~ 2016.03.31.
검증용 페이스북 텍스트 데이터	2,176	'남자들의 동영상' 페이지 2016.04.01. ~ 2016.05.31.

전처리 (인스타그램)

1. 문자열 기준 15이하 제거
2. 중복제거
3. 단어 가중치 기준 사용 형태소, Uni-gram 기반 변수 선택

전처리 (인스타그램, 페이스북)

1. NA값 제거
2. 트위터(Twitter)형태소 분류기를 사용한 형태소 분할
3. 지도학습방식의 단어 가중치 할당

2 모델링

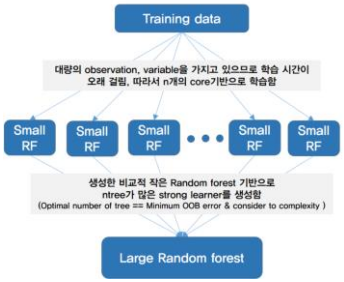
인스타그램 기반 랜덤포레스트

- 사용변수 (총 4,002개)
  1. 단어가중치 기준 4,000개 단어
  2. URL의 개수
  3. E-mail의 개수
- 파라미터
  1. 변수 개수 -> 200개
  2. 나무의 수 -> 65개

인스타그램 + 페이스북 기반 랜덤포레스트

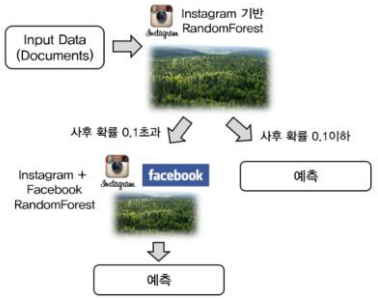
- 사용변수 (총 11개)
  인스타그램 랜덤포레스트 사후확률 포함 페이스북 특성을 반영한 파생 변수 11가지
- 파라미터
  1. 변수 개수 -> 4개
  2. 나무의 수 -> 100개

랜덤포레스트 분류기 사용



3 최종 모델 구축 및 평가 데이터 검증

통계적 검증기반(T-test) 최종 분류기 비교



데이터 특성	Accuracy	F1-measure
인스타그램 RF	0.944	0.680
페이스북 파생변수 RF	0.980	0.884
인스타그램 + 페이스북 파생변수 RF	0.993	0.960

평가 데이터(시간 기준 미래 데이터) 검증

페이지 이름	데이터 수	개시 기간
남자들의 동영상	2,176건	2016.04.01 ~ 2016.05.31

1. Literature review

- **한글 텍스트 기반의 비교적 대량의 SNS 스팸 필터링 알고리즘에 대한 연구가 거의 없는 실정임**

〈 해외 스팸 필터링 연구 〉

저자	데이터
Zhang et al. (2016)	1,181,735개의 트위터 계정을 기반으로 URL의 정보에 근거하여 트위터 계정간의 유사도를 측정하는 방법론과 그래프 기반 접근법을 제시한 연구
Yang et al. (2013)	14,400,000개의 트위터 계정을 사용 하여 그래프 기반 속성, 이웃 기반 속성, 시간 기반 속성, 자동화 기반 속성 등 4가지의 접근법을 제시한 연구
Gao et al.(2012)	187,000,000개의 페이스북 포스트와 17,000,000개의 트위터 트윗을 사용하여 개별 스팸을 탐지하는 것이 아닌 스팸 유포의 진원지를 찾는 연구
Yang et al., (2011)	14,401,157개의 트윗과 485,721개의 트위터 계정으로 트위터에서 스팸을 유포하는 스패머 계정을 찾는 연구
Stringhini et al. (2010)	900개의 Facebook, Myspace, Twitter 계정을 개설하여 SNS 친구 신청 건수, 메시지 건수, 방문 건수 등의 로그정보를 이용하여 스팸성 계정을 탐색한 연구
Kanaris et al. (2006)	2,893개의 이메일 데이터를 이용하여 BoW 기반 SVM을 사용하여 스팸 필터링 알고리듬을 구현한 연구

〈 국내 스팸 필터링 연구 〉

저자	데이터
이승제 외 (2011)	2,548 SMS 데이터를 기반으로 BoW기반의 Naïve bayes classifier, Support vector machine를 사용한 연구
이하나 외(2011)	500 SMS 데이터를 기반으로 여섯가지 절차의 rule based 알고리듬을 사용한 연구
이성욱 (2010)	9,740 E-mail 데이터를 기반으로 Chi-sq 가중치 기반의 Support vector machine을 사용한 연구
조인휘 외 (2009)	460 SMS 데이터를 기반으로 BoW 기반의 Support vector machine

〈한글 텍스트 기반 스팸 필터링 연구 현황〉

1. 소량의 이메일과 SMS 데이터 실험
2. 대량의 SNS 스팸탐지를 위하여 텍스트 기반의 스팸필터링 알고리즘을 위한 연구는 아직 없는 실정임

대량의 SNS기반 한글 텍스트 스팸 필터링 연구 필요

1908

2. Data collection

- 대량의 데이터를 수집 하는 것은 가능하지만, **'목적변수를 어떻게 정의 할 것인가?'** 라는 것이 하나의 문제임
- 회사 관계자의 경우는 집단지성을 이용할 수 있지만 학생으로써 불가능, 직접 하기에는 데이터가 너무 많음
- 초기 모델링에서 사용할 데이터는 Instagram기반으로 사용함 (과연 Facebook target을 대체할 수 있을까?)

facebook

Crawling Text & Image

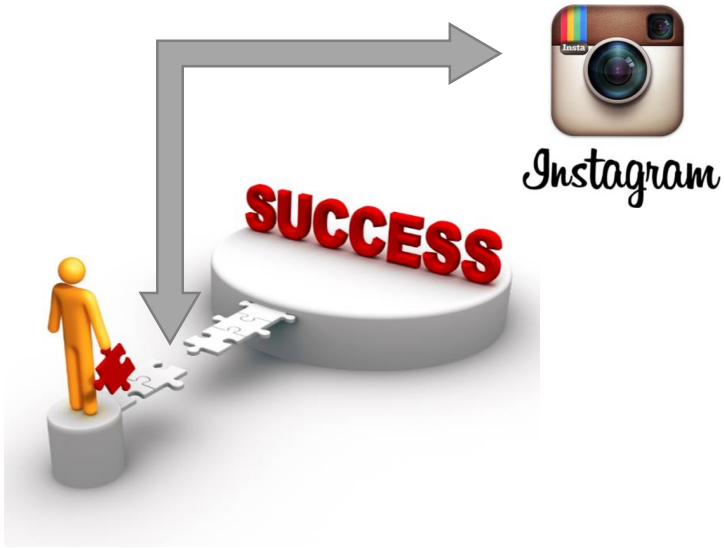
Input variable V1, V2, V3 … $V_{(n)}$ (Bag of Words)	Target (Ham/Spam)
수집가능 :)	수집불가능 :<
수집가능 :)	수집불가능 :<
수집가능 :)	수집불가능 :<
수집가능 :)	수집불가능 :<
수집가능 :)	수집불가능 :<

Solution

(1) 페이지 운영자라면 신고정보를 이용하여 Labeling 후 예측
(우리는 운영자도 아니고..) :<

(2) 직접 보면서 전부 Labeling
(비효율적인 엄청난 시간 소요, 실험비는 큰 돈임) ;< :<

(3) 다른 SNS데이터에서 training하여 예측해보자 :)
(Instagram의 Hash tag(#) 을 이용해 보자!)



1909

2. Data collection

- SPAM hash tag는 다양한 주제를 고려하여 총 70가지 선정, 초기 데이터 수는 287,140개임
- Hash tag가 중요한 정보이나, 공적인 장소에서 사용하기에는 부적절 하므로 숨김 처리 하였음

Hash Tag	데이터 수
[redacted]	2,464
[redacted]	1,966
[redacted]	5,558
[redacted]	1,955
[redacted]	1,551
[redacted]	2,000
[redacted]	472
[redacted]	2,000
[redacted]	4,000
[redacted]	6,975
카지노사이트	6,000
[redacted]	2,878
[redacted]	1,350
[redacted]	3,135
[redacted]	4,915
[redacted]	2,347
[redacted]	3,980
[redacted]	3,923
[redacted]	3,466
[redacted]	2,000
[redacted]	1,998
[redacted]	5,470
라이브바카라	5,861
[redacted]	7,587

Hash Tag	데이터 수
[redacted]	8,913
[redacted]	5,755
[redacted]	769
[redacted]	3,021
[redacted]	4,654
[redacted]	4,938
[redacted]	2,972
[redacted]	2,766
[redacted]	5,563
[redacted]	1,718
소액대출	1,299
스포츠베팅	3,992
[redacted]	4,854
[redacted]	9,020
[redacted]	4,917
[redacted]	277
[redacted]	1,873
[redacted]	7,575
[redacted]	4,000
[redacted]	1,734
[redacted]	5,466
[redacted]	6,980
라이브스코어	2,000

Hash Tag	데이터 수
대출상품	253
메이저놀이터	2,762
[redacted]	8,976
[redacted]	4,311
[redacted]	7,829
[redacted]	2,948
네임드사다리분석	3,400
[redacted]	7,185
[redacted]	6,243
[redacted]	6,097
[redacted]	3,756
[redacted]	1,091
[redacted]	6,905
[redacted]	4,995
[redacted]	4,798
[redacted]	5,927
[redacted]	1,977
[redacted]	5,591
[redacted]	2,031
[redacted]	4,931
안전한놀이터	3,960
[redacted]	5,909
[redacted]	6,358

1910

2. Data collection

HAM ☺

- HAM Hash Tag는 총 81개, 초기 데이터 수는 455,679개임

Hash Tag	데이터 수
짜장면	4,820
책	4,980
브런치	7,980
석가탄신일	7,000
이별	6,980
카페	6,977
캣스타그램	3,200
춘천	5,000
커피스타그램	4,635
소통	7,712
크루저보드	2,000
데일리	8,706
일상스타그램	7,000
인연	4,999
두산베어스	5,000
먹스타그램	7,800
감성	8,985
감성사진	6,000
공감	4,160
운동	6,824
얼스타그램	6,906
가족사진	4,000
축제	1,620
선팔	7,951
팔로미	6,000
친스타그램	5,970
좋은글	5,996
글귀	7,997

Hash Tag	데이터 수
행복	6,993
하리보	3,440
힐링	7,987
운동하는남자	5,599
운동하는여자	6,998
안녕	7,000
취미	6,000
주부	5,000
과제	3,620
허니버터칩	1,400
아이스크림	7,980
인스타베이비	4,980
맞팔	7,703
잔치국수	3,980
김치전	4,000
도서관	6,000
일상생활	8,745
좋아요	7,942
럽스타그램	6,961
사랑해	5,000
결혼	7,000
추억	6,952
동기유발	4,000
영화스타그램	3,080
남산	5,000
올림픽공원	5,000
소풍	5,000
손글씨	6,994

Hash Tag	데이터 수
피크닉	5,958
먹스타그램	7,843
족발	6,460
선물	8,000
근황	3,600
로즈데이	8,000
셀피	9,854
학생	5,999
서핑	4,000
티타임	2,000
시험기간	2,980
글귀스타그램	7,000
글스타그램	4,078
만리장성	600
티라미수	3,000
나무	7,800
여행	7,948
여행스타그램	6,974
진심	2,440
산책	4,986
날씨	6,994
주말	3,913
명언	2,500
세계여행	2,000
짬뽕	4,000
손글씨	6,994

1911

2. Data collection

- Facebook 데이터는 총 23개의 페이지에서, 1,795,067개의 데이터를 습득함

1,795,067개의 데이터

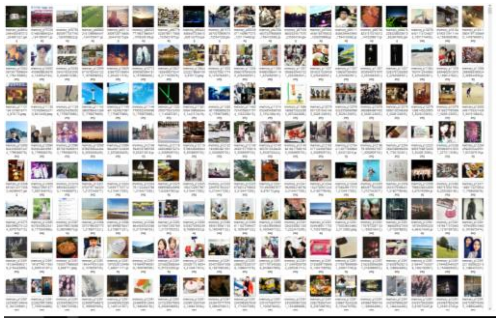
페이지 이름
세상에서 가장 웃긴 동영상
남자들의 동영상
여행에 미치다
네임드사다리
남자의 관리
남자옷덕후
여자들의 동영상
메갈리아4
레진코믹스
잡지 사심
한국성폭력상담소
민주노총
스포츠 마니아 일루모여
대학내일
남자들의 축구
세상에서 가장 소름돋는 라이브
19세 이상만
중고차 전국출장매입
도우미론
사다리전문페이지
안전놀이터/네임드사다리 ｂｓ－ｓｋｙ－１．ｃｏｍ 안전토토
바카라 놀이터 네임드사다리
사다리프젝전문
세상에서 가장 웃긴 동영상
남자들의 동영상
여행에 미치다

1912

3. Data preprocessing

- Instagram에서 크롤링한 데이터를 바탕으로 1차 전처리함
- SPAM의 경우 약 2만건 이므로, HAM이 섞여 있을 경우 전수조사를 통해 목표변수를 수정함

	A	B	C	D	E	F
1		image_source	imagepost_link	data_name	image_id	title
2	1	https://scontent.cdninstagram.cc	https://www.instagram.com/p/BE	Mino. 24. 부산.	p1238748349838230610_2811673424	저도 득자라며 징징거렸더니 #시 한편을 써 주신#조이글 님 #취향저격 입니다~!! 어제 받았는데 이제와서 자랑입니다ㅋㅋ - - #Mino울림#시#자작시#단편시#시스타그램#글쟁이#글#글귀#좋은글#글스타그램#감성#감스타그램#공감#공스타그램#명언#사랑#사랑글#짝사랑#이별#이별글#연인#연애#데일리#일상#하루#힐링#소통#좋아요



Instagram
HAM 81가지
SPAM 70가지 Hashtag기반
742,819개 텍스트 & Image 수집

spam		ham	
eng	kor	eng	kor
		hamtext	글귀
		life	일상
		self	셀피
		emotion	감성
		handwrite	손글씨
		daily	데일리
		pig	먹스타그램
		interactiveFollow	맞팔
		like	좋아요
		bye	이별
		trip	여행
		text	글귀스타그램
		excercise	운동
		healing	힐링
		communication	소통
		friendStagram	친스타그램
		destiny	인연
		happiness	행복
		goodtext	좋은글
		loveStagram	럽스타그램
		marry	결혼
		firstFollow	선팔
		faceStagram	얼스타그램
		memory	추억
		healthyWoman	운동하는여자
		healthyMan	운동하는남자
		tripStagram	여행스타그램
		movieStagram	영화스타그램
		coffeeStagram	커피스타그램
		termexam	시험기간
		homework	과제
		weather	날씨
		picnic	소풍
		walk	산책
		recent	근황
		festival	축제
		TRUE	인싸
		family	가족사진
		weekend	주말
		wiseSaying	명언
		textStagram	글스타그램
		student	학생
		library	도서관
		cafe	카페

HAM	SPAM
455,679 records	287,140 records

1차 전처리 목록

1. nchar기준 15이하 제거
2. NA 값 제거
3. 중복 제거

HAM	SPAM
178,765 records ➔ 181,590 records (수작업..)	20,530 records ➔ 17,705 records (수작업..)

1913

3. Data preprocessing

- 2차 전처리로 Twitter기반의 POS tagger를 사용하여 시행함
- Supervised term weighting 방법중 DF기준으로 Prob_W와 Entropy기반의 Information Gain을 사용함
- Twitter POS tagger는 TagSet중 Email, URL을 지원하기 때문에 3가지에 대한 Meta 변수를 추출함

POS tagging
'KoNLPy'
(Twitter POS Tagger)

Supervised term weighting method based on contingency table

Feature \ Category	c_j	$\overline{c_j}$
f_i	$f_i c_j$ (1)	$f_i \overline{c_j}$ (2)
$\overline{f_i}$	$\overline{f_i} c_j$ (3)	$\overline{f_i} \overline{c_j}$ (4)

(1) $f_i c_j$ → c_j클래스에서 f_i term이 있는 document 수
(2) $f_i \overline{c_j}$ → c_j가 아닌 클래스에서 f_i term이 있는 document 수
(3) $\overline{f_i} c_j$ → c_j클래스에서 f_i term이 없는 document 수
(4) $\overline{f_i} \overline{c_j}$ → c_j가 아닌 클래스에서 f_i term이 없는 document 수

따라서 (1)+(2)+(3)+(4) = 전체 document 수

*안녕하세요! 제카카오톡아이디는dsba5207이고, 사이트 주소는 http://dsba.korea.ac.kr/wp/ 입니다.

패키지	결과
Kkma (KoNLPy)	[(*, SW), (안녕하, VA), (세요, EFN), (!, SF), (저, NP), (의, JKG), (카카오, NNG), (톡, MAG), (아이, NNG), (디, NNG), (는, JX), (dsba, OL), (5207, NR), (이, VCP), (고, ECE), (., SP), (사이트, NNG), (주소, NG), (는, JX), (http, OL), (:, SP), (//, SW), (dsba, OL), (., SF), (korea, OL), (., SF), (ac, OL), (., SF), (kr, OL), (/, SP), (wp, OL), (/, SP), (이, VCP), (ㅂ니다, EFN), (., SF)]
Twitter (KoNLPy)	*(Punctuation: 0, 1), 안녕하다(Adjective: 1, 5), !(Punctuation: 6, 1), 제(Noun: 8, 1), 카카오(Noun: 9, 3), 톡(Noun: 12, 1), 아이디(Noun: 13, 3), 는(Josa: 16, 1), dsba(Alpha: 17, 4), 5207(Number: 21, 4), 이고(Josa: 25, 2), .(Punctuation: 27, 1), 사이트(Noun: 29, 3), 주소(Noun: 33, 2), 는(Josa: 35, 1), http://dsba.korea.ac.kr/wp/(URL: 37, 27), 이다(Adjective: 65, 3), .(Punctuation: 68, 1)

Weight method
1. Prob Weight (one class)
2. Information Gain (Entropy) (two class)

$$W(f) = tf * \log\left(1 + \frac{fc}{\overline{fc}} \frac{fc}{\overline{fc}}\right)$$

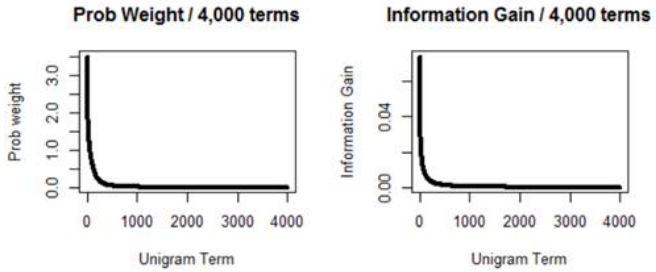
TF-PROB equation

1914

3. Data preprocessing

- Instagram에서 생성된 전체 Term은 144,187개임, 짧은 문서의 특성상 sparse함
- 모든 Term를 사용 하는 것은 PC 환경상 비효율적이며, 성능도 좋지 않음
- Prob weight, Information Gain기준 상위 4,000개의 term만 사용함

Term	Spam_TF	Ham_TF	Total_TF	Spam_DF	Ham_DF	CM_1_1	CM_1_2	CM_2_1	CM_2_2	Chi_sq_W	Prob_W
☺	1813	0	1813	1787	0	0	1787	182337	18878	0.0000534026943724	5.13778839825791
	2724	2	2726	2508	2	2	2508	182335	18157	0.0003176858455420	4.75841648707327
	3516	6	3522	2726	5	5	2726	182332	17939	0.0011412944718870	4.24973790011247
	1713	2	1715	1507	2	2	1507	182335	19158	0.0005853681429181	3.70289983355734
	1430	0	1430	863	0	0	863	182337	19802	0.0001211131133584	3.65573899162882
	9328	45	9373	3919	29	29	3919	182308	16746	0.0173044818178061	3.45269430704749
	9095	33	9128	3333	20	20	3333	182317	17332	0.0106629033522763	3.45118774670497
	11558	95	11653	6115	87	87	6115	182250	14550	0.0723742825739593	3.40823507465256
	3080	2	3082	1019	1	1	1019	182336	19646	0.0004037659983303	3.31336111117449
	16375	324	16699	4262	50	50	4262	182287	16403	0.0440124164318217	3.1233576540734
	23078	486	23564	4846	72	72	4846	182265	15819	0.0737582932712828	3.06072755039345
	23071	449	23520	5856	144	144	5856	182193	14809	0.210191107207184	2.83171246414486
	5219	76	5295	3584	48	48	3584	182289	17081	0.052125957834961	2.79452004528266
	3668	45	3713	3263	39	39	3263	182298	17402	0.0395916125929114	2.79143475205842
	10682	129	10811	4686	96	96	4686	182241	15979	0.136562065511234	2.71947367863914
	2711	5	2716	1275	5	5	1275	182332	19390	0.0028244232668767	2.70766736148212



사용한 형태소 모양

Adjective
Adverb
Alpha
Foreign
Koreanparticle
NA
Noun
Verb

Prob Term + URL + Email

Prob Term(4,000) + URL (1) + Email (1)
총 4,002개의 Input variable 생성

IF Term + URL + Email

IF Term(4,000) + URL (1) + Email (1)
총 4,002개의 Input variable 생성

Instagram classifier 생성

1915

4. Modeling

– Text classifier

- RandomForest는 unbalanced data에 민감한 특성과 본 연구에서 사용할 경우 PC에서는 비교적 대량의 메모리를 필요로 함
- Down sampling을 하여 모델링을 하였음 (HAM 25,000개, SPAM 17,705개)
- SPAM Term을 기준으로 가져온 Prob_W방식을 채택함 (One class)
- 변수의 수는 보통 sqrt(4,002), 약 63개를 사용하지만, sparse한 구조이기 때문에 200개를 사용함

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

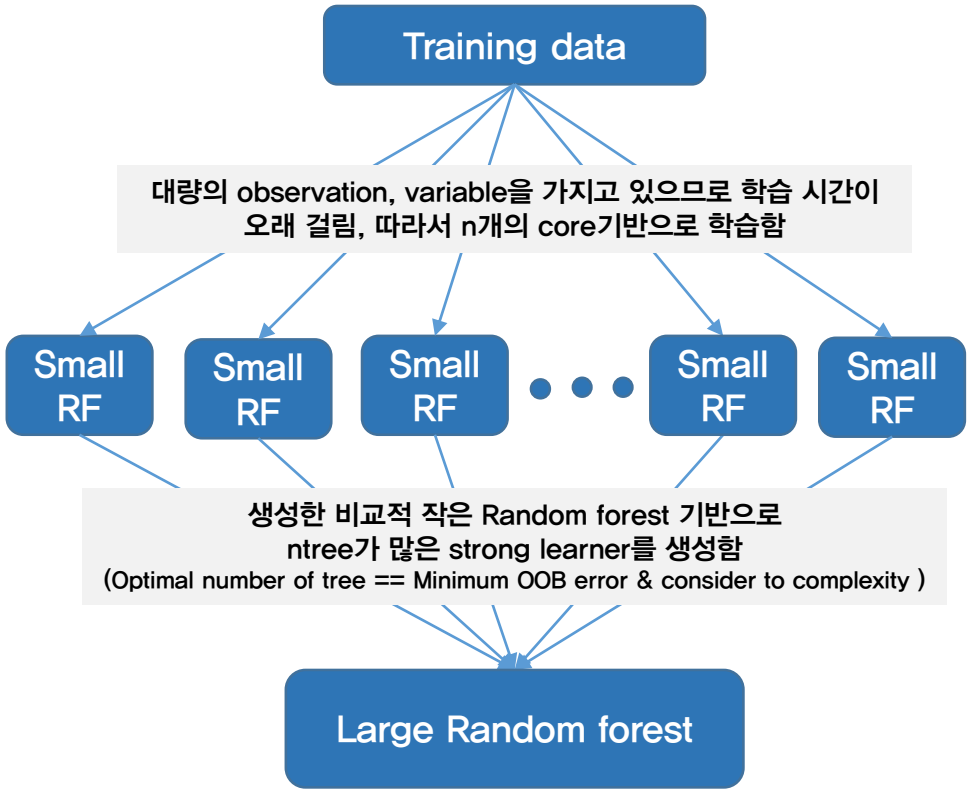
Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, Dinani Amorim 15(Oct):3133–3181, 2014.

Abstract

We evaluate **179 classifiers** arising from **17 families** (discriminant analysis, Bayesian, neural networks, support vector machines, decision trees, rule-based classifiers, boosting, bagging, stacking, random forests and other ensembles, generalized linear models, nearest-neighbors, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and other methods), implemented in Weka, R (with and without the caret package), C and Matlab, including all the relevant classifiers available today. We use **121 data sets**, which represent **the whole UCI** data base (excluding the large-scale problems) and other own real problems, in order to achieve significant conclusions about the classifier behavior, not dependent on the data set collection. **The classifiers most likely to be the bests are the random forest** (RF) versions, the best of which (implemented in R and accessed via caret) achieves 94.1% of the maximum accuracy overcoming 90% in the 84.3% of the data sets. However, the difference is not statistically significant with the second best, the SVM with Gaussian kernel implemented in C using LibSVM, which achieves 92.3% of the maximum accuracy. A few models are clearly better than the remaining ones: random forest, SVM with Gaussian and polynomial kernels, extreme learning machine with Gaussian kernel, C5.0 and avNNet (a committee of multi-layer perceptrons implemented in R with the caret package). The random forest is clearly the best family of classifiers (3 out of 5 bests classifiers are RF), followed by SVM (4 classifiers in the top-10), neural networks and boosting ensembles (5 and 3 members in the top-20, respectively).

[abs][pdf][bib]

behavior, not dependent on the data set collection. **The classifiers most likely to be the bests are the random forest** (RF) versions, the best of which



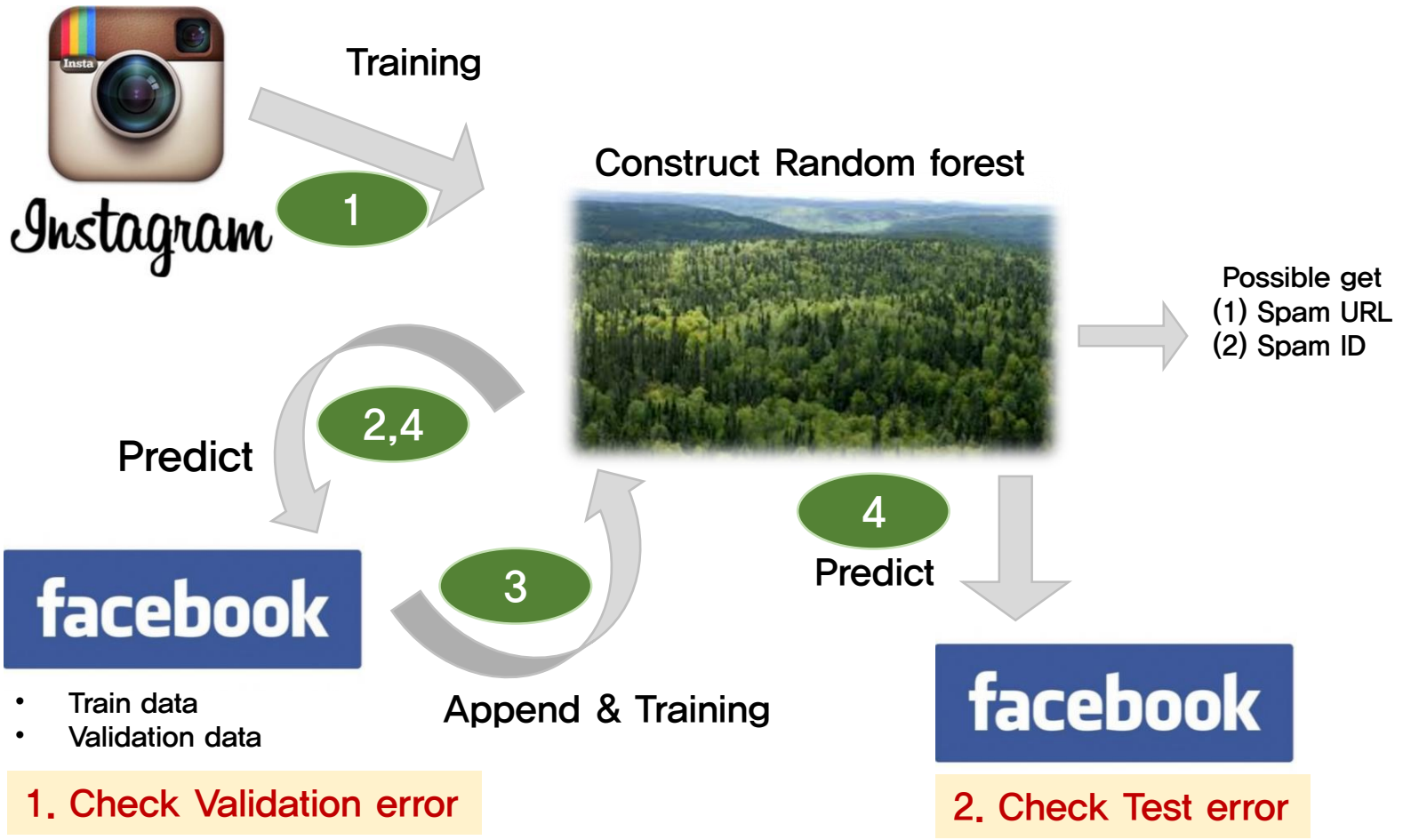
1916

Fernández-Delgado. M., Cernadas. E. (2014) ***Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?*** Journal of Machine Learning Research, 15, 3133-3181

4. Modeling

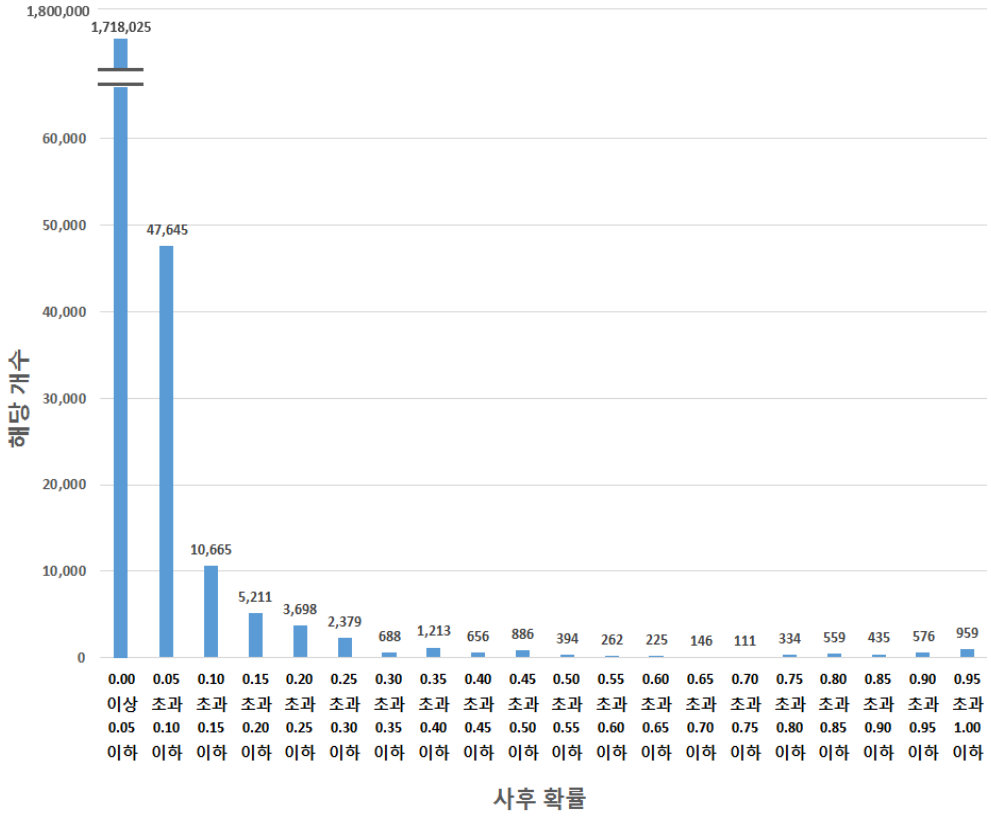
– Text classifier

- 부가적으로 본 프레임워크에서 SPAM URL, SPAM ID를 생성할 수 있음 (Rule base 생성 가능)



4. Modeling
– Text classifier

- Instagram의 4,002개의 변수로 생성한 Random forest를 사용하여 분류함
- Facebook 데이터는 총 1,795,067개임
- 0.1 이하로 1,749,428개이며, 0.1초과로 29,397개임
- 스팸 확률 0.1 이하의 문서를 무작위 추출하여 확인한 결과, 스팸성 게시글은 한 건도 존재하지 않음 ➔ 이 경우 Ham으로 가정함



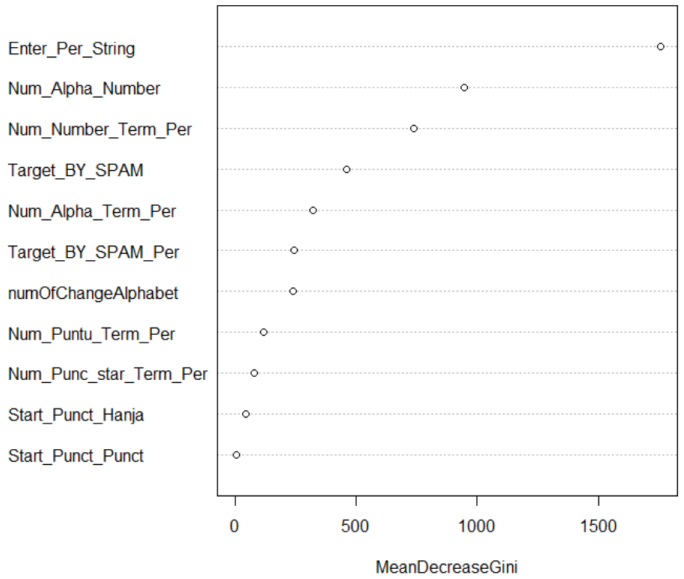
RandomForest posterior based on Facebook documents

오분류 예시 (Type2 Error)	오분류 예시 (Type1 Error)	정분류 예시
도박 마약ㅌㅋㅋㅋㅋㅋㅋ 짱웃갸 (1.000)	2개 무료 증 정 분석깅 PD80 (0.123)	매일 단톡방에서 진행되는 사다리프젝. 인원이다차면 모집을 안합니다. 선착순30명만 받습니다. 100%무료픽 단톡방입니다. 카톡 star345 (0.831)
대출해서 밥사줘야할거같다 (0.923)	안녕하세요^^ :) 페이스북 자동 댓글 프로그램 임대!! # 시간당 100개이상# 다중 아이디, 이미지업로드 가능 카톡 sunharu 연락주세요. (0.307)	♥http://me2.do/G3bcSoyE 〈--채팅계의 거물 20~40대 여성,남성 분포 데이트,애인대행,하룻밤 그냥 무료로 즐기실분만 가입하세요♥ (0.615)
그러나 따봉충 토토충은 극혐... (0.846)	23살 지연이에여. 카톡:gae22 추가! (0.246)	신용 믿음 100%믿을수잇는 출장대행 홈피주소 ｗｗｗ．ｋｉＳＳ３４３．Ｃｏｍ 섹시한 여대생 대기중입니다 (0.723)

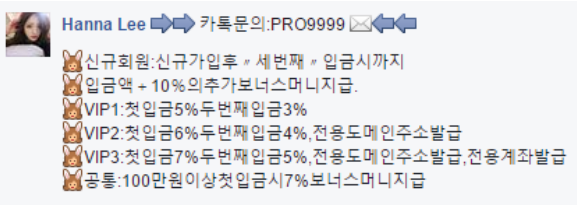
4. Modeling
– Text classifier

- Facebook의 스팸문서탐지에 유용한 총 10가지의 파생 변수를 생성함
- 파생변수와 Instagram기반 사후 확률을 다시 입력변수로 사용하여 RandomForest 생성함 nTree=100, nVar=5 설정

변수명	설명
Enter_Per_String	엔터의수 / 스트링길이
Num_Alpha_Number	알파벳 + 숫자 Term 개수
Num_Number_Term_Per	숫자 Term 개수 / 스트링 길이
Target_BY_SPAM	Instagram RandomForest Probability
Num_Alpha_Term_Per	알파벳 Term 개수 / 스트링 길이
Target_BY_SPAM_Per	Instagram RandomForest Probability / 스트링 길이
Num_of_Change_Alphabet	대소문자 알파벳 Term 개수
Num_Puntu_Term_Per	한자기반 특수문자 개수 / 스트링 길이
Num_Punc_star_Term_Per	쉬프트 기반 특수문자 개수 / 스트링 길이
Start_Punct_Hanja	한자기반 특수문자로 시작유무
Start_Punct_Punct	쉬프트기반 특수문자로 시작유무



Facebook Spam의 특징이 있음



4. Modeling

– Text classifier

- 모든 평가 지표에서 Instagram RF결과만 사용한 것보다 Instagram결과 + Facebook변수로 만든 RF가 더 좋은 성능을 산출함
- 6개의 평가지표 모두 T-test p-value < 0.01 로 산출되며 결국 통계적으로 유의미함
- 본 결과는 Testdata를 2000개로 설정하고 50번 반복 실험 한 것에 대한 결과임

Model / Valid Index	인스타그램 RF (Model A)	페이스북 파생변수 RF (Model B)	인스타그램 + 페이스북 파생변수 RF (Model C)
Sensitivity	0.728(0.037)	0.953(0.017)	**0.963(0.015) (**,**)**
Precision	0.641(0.038)	0.824(0.027)	**0.958(0.014) (**,**)**
Specificity	0.963(0.004)	0.982(0.003)	**0.996(0.001) (**,**)**
Accuracy	0.944(0.005)	0.98(0.003)	**0.993(0.002) (**,**)**
BCR	0.837(0.021)	0.967(0.008)	**0.979(0.008) (**,**)**
F1	0.68(0.029)	0.884(0.018)	**0.960(0.009) (**,**)**

Mean (Standard deviation)
** ➔ p-value < 0.01

전 부분 T-test p-value 0.01보다 적으며, 통계적으로 유의미함 ☺

Hypothesis 1

$$\mu_{Valid\ Index(i)}^{Model\ 1} = \mu_{Valid\ Index(i)}^{Model3}$$

versus

$$\mu_{Valid\ Index(i)}^{Model\ 1} < \mu_{Valid\ Index(i)}^{Model\ 3}$$

Hypothesis 2

$$\mu_{Valid\ Index(i)}^{Model\ 2} = \mu_{Valid\ Index(i)}^{Model\ 3}$$

versus

$$\mu_{Valid\ Index(i)}^{Model\ 2} < \mu_{Valid\ Index(i)}^{Model\ 3}$$

4. Modeling

– Text classifier

- 반복 실험 후 생성된 분류기로, 생성된 모델의 데이터 기준 미래 데이터(test data)로 실제 적용해 보았음
- 실제 남자들의 동영상 페이지에서 수집 하여 이를 실험해봄 (2016.04.01 ~ 2016.05.31)
- 다른 페이지에 비하여 비교적 스팸 비중이 높아 보였기 때문에 선정함

(2016.04.01 ~ 2016.05.31)
총 2,176개의 포스팅,댓글 분류

분류기도 잘 만들어 졌으니 보기 싫었던
최근 스팸 댓글, 포스팅 걸러 보겠습니다

1921

4. Modeling

– Text classifier

- 결론적으로 2,176개 중 1개를 오분류함
- 기존 Cross Validation 결과가 크게 차이가 나지 않는 것을 확인 할 수 있음
- 가변적인 길이에도 분류 결과가 우수하게 나타나는 것을 알 수 있음

(2016.04.01 ~ 2016.05.31)
총 2,176개의 포스팅,댓글 분류

Confusion Matrix		Actual	
		SPAM	HAM
Predict	SPAM	27	0
	HAM	1	2,148

Valid Index	Sensitivity	Precision	Specificity	Accuracy	BCR	F1
Results	1.000	1.000	0.964	0.999	0.982	0.982

1922

5. Conclusion

가설 1. Instagram에서 생성한 Transfer learning 모델은 Facebook spam 분류가 잘 작동할 것이다.

➔ 아주 잘 작동하지는 않음

가설 2. Instagram 분류기와 함께 Facebook에서 생성한 meta 변수를 통해 만든 분류기는 잘 작동할 것이다.

➔ Accuracy기준 99% 이상이며 F1-measure 기준으로 약 96%의 분류기를 생성할 수 있음

➔ Instagram 분류기에 비하여 F1, Accuracy, Recall, Sensitivity, Specificity 모든 지표에서 통계적으로 유의미함

기여점

➔ 스팸 필터링 주제에서 만큼은 기존 데이터 수집 후 Supervised Learning의 문제점인 target을 대체하는 방식으로 Instagram에서 학습하는 방법이 적합 한 것을 증명함

➔ 따라서, 데이터의 취득권을 가진 특수 계층만이 아닌 일반 사용자도 오픈소스로 가용할만한 SNS 텍스트 스팸 필터링을 생성할 수 있는 프레임워크임

➔ SNS에서의 문제점 중 하나인 가변적인 길이(짧은)의 스팸 처리도 Meta변수를 통해 극복할 수 있다는 것을 확인함

6. Future work

향후과제 1. Facebook의 대량의 데이터를 적재 후, 페이스북 단어 기반 스팸필터링을 통하여 성능검증

➔ 본 연구는 데이터의 목표변수를 가용할 수 있는 인스타그램의 학습된 단어로 1차적으로 페이스북 스팸 필터링을 검증

➔ 적재된 페이스북 데이터 단어기반의 분류기와 인스타그램과의 성능비교 연구필요

향후과제 2. 문법 파괴적인 SNS의 단어에 대한 원형 처리가 필요

➔ 단어 기반의 스팸 필터링 분류기가 잘 작동하지 않는 하나의 이유는 스팸 단어에 대한 회피가 발생하여 BoW기반의 한계점이 발생하기 때문임

향후과제 3. 새로운 스팸주제에 관한 키워드 탐지필요

➔ 제시한 프레임워크에서 초기조건으로 필요한 부분은 스팸에 대한 주제의 인식임. 따라서 대량의 데이터에서 스팸 주제에 관한 탐지에 대한 연구가 진행 되어야 함

Reference

[1] Breiman L. (2001). ***Random Forests.*** Machine Learning, 45(1), 5–32

[2] Fernández-Delgado. M., Cernadas. E. (2014) ***Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?***, Journal of Machine Learning Research, 15, 3133-3181

[3] Gao, H., Chen, Y., Lee, K., Palsetia, D., Choudhary, A., N. (2012). ***Towards Online Spam Filtering in Social Networks,*** In NDSS 12, 1-16

[4] Jo. C., Y. (2011) ***A Semiotic Study for New Media -applied to the case for Social Network Service,*** Semiotic Inquiry, 30, 125-154.

[5] Joe, I., H., Shim, H., T. (2009) ***A SVM-based Spam Filtering System for Short Message Service*****,** The Korean Institute of Communications and Information Sciences, 34(9), 908-913.

[6] Kanaris, I., Kanaris, K., Stamatatos E.(2006), ***Spam detection using character n-grams***, Hellenic conference on artificial intelligence. 3955, 95-104

[7] Lee, H., N., Song, M., G., Im., E., G. (2011) ***A Study on Structuring Spam Short Message Service(SMS) filter,*** The Korean Institute of Communications and Information Sciences, 1072-1073

[8] Lee, S., J., Choi, D., J. (2011) ***Personalized Mobile Junk Message Filtering System ,*** The Journal of the Korea Contents Association, 11(12), 122-135.

[9] Lee, S., W. (2010) ***Spam Filter by Using X2 Statistics and Support Vector Machines ,*** The KIPS transactions, 17(3), 249-254

[10] Quan. X., Liu. W. and Qiu. B. (2011),***Term Weighting Schemes for Question Categorization***, IEEE Transactions on Pattern Analysis and Machine Intelligence archive , 33(5), 1009-1021

[11] Soiraya, M., Thanalerdmongkol, S., & Chantrapornchai, C. (2012). ***Using a Data Mining Approach: Spam Detection on Facebook,*** International Journal of Computer Applications, 58(13).

[12] Stringhini, G., Kruegel, C., Vigna G. (2010), ***Detecting spammers on social networks***, Proceedings of the 26th Annual Computer Security Applications Conference, 1-9

[13] Yang, C., Harkreader, R. C., & Gu, G. (2011). ***Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers,*** In International Workshop on Recent Advances in Intrusion Detection, 318-337.

[14] Yang, C., Harkreader, R. C., & Gu, G. (2013), ***Empirical evaluation and new design for fighting evolving Twitter spammers***, IEEE Transactions on Information Forensics and Security, 8(8), 1280-1293.

[15] Zhang, X., Li, Z., Zhu, S., Liang, W. (2016), ***Detecting spam and promoting campaigns in Twitter,*** ACM Transactions on the Web (TWEB) 10(1)

[16] Zheng, X., Zeng, Z., Chen, Z., Yu, Y., & Rong, C. (2015). ***Detecting spammers on social networks,*** Neurocomputing, 159, 27-34.

1926

THANK YOU

이 논문은 2016년도 정부(미래창조과학부)의 재원으로
한국연구재단의 지원을 받아 수행됨(No. 2014R1A1A1004648, No. NRF-2015R1A2A2A04007359)