

Instagram, Facebook 데이터를 기반으로한 Facebook 스팸 필터링 알고리즘 개발

김준홍, 김해동, 서덕성

School of Industrial Management Engineering

Data Science & Business Analytics LAB

1. Purpose of the project

- 최근, 저비용으로 확산속도가 빠른 SNS기반의 스팸 광고가 활성화 되고 있음
- 페이스북의 유명 페이지의 게시물에 달리는 댓글과 사용자가 올린 게시물 중 스팸성 댓글과 게시물을 필터링 하고자 함
- 오른쪽 예시와 같이 text기반에서는 'Ham' 이지만 Image기반에서는 'Spam' 으로 분류되는 고도화된 게시물이 생겨나고 있음

Case1. Text Spam & Image Spam Case2. Text Spam



Case3. Text Ham, Image Spam



1. Purpose of the project

- 최근, 저비용으로 확산속도가 빠른 SNS기반의 스팸 광고가 활성화 되고 있음
- 페이스북의 유명 페이지의 포스트에 달리는 댓글과 사용자가 올린 포스트 중 스팸성 댓글과 포스트를 필터링 하고자 함
- 오른쪽 예시와 같이 text기반에서는 'Ham' 이지만 Image기반에서는 'Spam' 으로 분류되는 고도화된 게시물이 생겨나고 있음

Case1. Text Spam & Image Spam Case2. Text Spam



Case3. Text Ham, Image Spam



〈시작 동기〉

개인적으로 Facebook을
학술적으로, 유익한 글 등의
목적으로 사용하는데,
Spam들이 보기 좋지 않아
지워보고 싶었음

1. Purpose of the project

- 최근, 저비용으로 확산속도가 빠른 SNS기반의 스팸 광고가 활성화 되고 있음
- 페이스북의 유명 페이지의 포스트에 달리는 댓글과 사용자가 올린 포스트 중 스팸성 댓글과 포스트를 필터링 하고자 함
- 오른쪽 예시와 같이 text기반에서는 'Ham' 이지만 Image기반에서는 'Spam' 으로 분류되는 고도화된 게시물이 생겨나고 있음

Case1. Text Spam & Image Spam Case2. Text Spam



Case3. Text Ham, Image Spam



<시작 동기>

개인적으로 Facebook을
학술적으로, 유익한 글 등의
목적으로 사용하는데,
Spam들이 보기 좋지 않아
지워보고 싶었음



<드릴 말씀..>

연구실 프로젝트 아닙니다..T
순전히 해보고 싶어서 시작한
거라는걸 알아 주셨으면..

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 label을 대체할 수 있을까?)

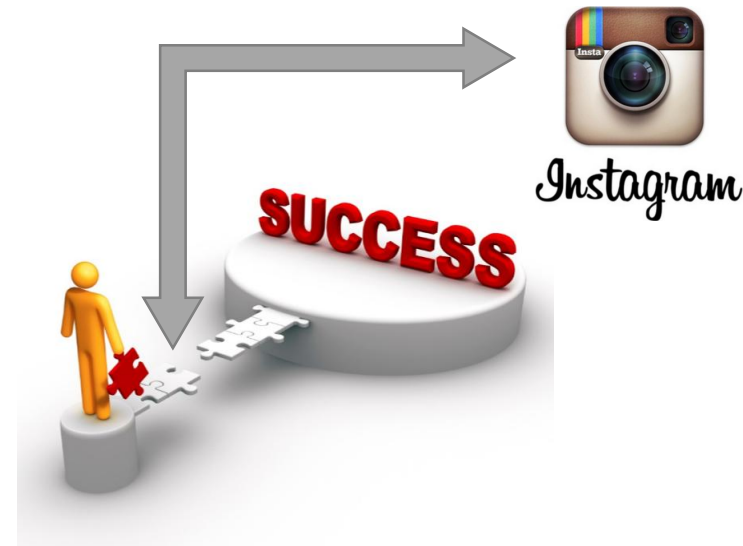


Crawling Text & Image

Input variable V1, V2, V3 ... V _(n) (BOW)	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!)



2. Data collection

SPAM :<

- SPAM Hash Tag는 총 70개, 초기 데이터 수는 287,140개임

우리말	데이터 수
텐프로	2,464
애인대행	1,966
성인	5,558
섹파구함	1,955
바카라	1,551
배팅	2,000
숨가그램	472
부산유흥	2,000
출장샵	4,000
카지노	6,975
카지노사이트	6,000
먹튀검증	2,878
시알리스	1,350
조건만남	3,135
일탈그램	4,915
은꼴	2,347
풀살롱	3,980
도박	3,923
강남풀싸롱	3,466
출장마사지	2,000
급전	1,998
스포츠토토잘하는법	5,470
라이브바카라	5,861
라이브카지노	7,587

우리말	데이터 수
대출상품	253
메이저놀이터	2,762
만남	8,976
신음소리	4,311
네임드	7,829
네임드 사다리	2,948
네임드사다리분석	3,400
오피녀	7,185
오피남	6,243
온라인바카라	6,097
오피걸	3,756
내구제	1,091
픽스터	6,905
사설놀이터	4,995
프로토	4,798
안전놀이터추천	5,927
룸사롱	1,977
사다리픽	5,591
사다리사이트	2,031
안전공원	4,931
안전한놀이터	3,960
안전사이트	5,909
섹텍	6,358

우리말	데이터 수
섹스	8,913
섹코드	5,755
물뽕파는곳	769
섹꼬드	3,021
섹그램	4,654
섹스파트너	4,938
노예녀	2,972
섹택	2,766
섹스타그램야스타그램	5,563
성인기구	1,718
소액대출	1,299
스포츠베팅	3,992
스포츠베팅사이트	4,854
스포츠토토	9,020
섯다	4,917
흥분제	277
선릉안마	1,873
토토	7,575
토토사이트	4,000
비아그라	1,734
여자신음	5,466
야동	6,980
라이브스코어	2,000

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
맛팔	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

2. Data collection

- Facebook 데이터는 총 23개의 페이지에서, 1,795,067개의 데이터를 습득함 (궁극적으로 예측할 데이터)



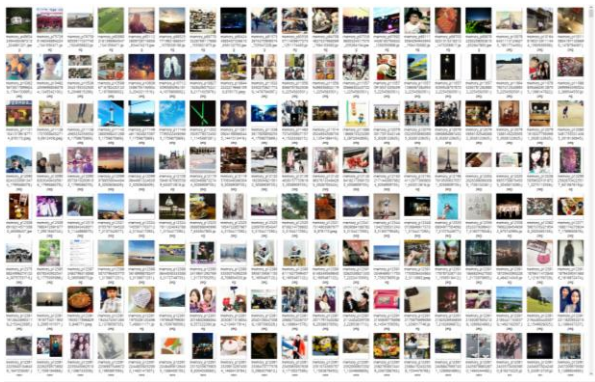
1,795,067개의 데이터

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 b s - s k y - 1 . c o m 안전토토
바카라 놀이터 네임드사다리
사다리프젝전문
세상에서 가장 웃긴 동영상
남자들의 동영상
여행에 미치다

3. Data preprocessing

- Instagram에서 크롤링한 데이터를 바탕으로 1차 전처리함
- SPAM의 경우 약 2만건 이므로, HAM이 섞여 있을 경우 '노동의 시간' 을 가짐

	A	B	C	D	E	F
1		image_source	imagepost_link	data_name	image_id	title
1		https://scontent.cdninstagram.com/p/BEMino.24.부산	https://www.instagram.com/p/BEMino.24.부산	p1238748349838230610_2811673424		저도 독자라며 장정거렸더니 #시 한편을 써 주신#조이글 남 #취향저격 입니다!! 어제 받았는데 이제와서 자랑입니다ㅋ - #Mino올림#시#자작시#단편시#시스타그램#글쟁이#글#글귀#좋은글#글스타그램#감성#감스타그램#공감#공스타그램#명언#사랑#사랑글#작사랑#이별#이별글#연인#연애#데일리#일상#하루#일링#소통#좋아요



	A	B	C	D
1	spam		ham	
2	eng	kor	eng	kor
3	meeting	만남	hamtext	글귀
4	adultery	성인	life	일상
5	deviate	일탈그림	well	잘피
6	opgit	오픈그림	emotion	감성
7	offgit	오프그림	handwrite	손글씨
8	offman	오프남	daily	데일리
9	sex.slavery	노예녀	pig	돼스타그램
10	boobs	송가그림	interactivefollow	맞팔
11	roomsalon	룸사롱	like	좋아요
12	gangnam.salon	강남룸싸롱	bye	이별
13	full.salon	풀룸싸롱	trip	여행
14	sunhill.massage	선돌안마	text	글스타그램
15	busan	부산유희	exercise	운동
16	home.massage	홈안마사지	healing	힐링
17	ask.fuckbuddy	섹파구함	communication	소통
18	acting_lover	액팅대행	friendstagram	친스타그램
19	sex	섹스	destiny	연연
20	moaning	신음소리	happiness	행복
21	sexgram	섹그림	goodtext	좋은글
22	sportstoto	스포츠토토	lovestagram	원스타그램
23	toto	토토	marry	결혼
24	gambling	도박	firstfollow	선팔
25	baccarat	바카라	facestagram	얼스타그램
26	livescore	라이브스코어	memory	추억
27	stimulant	흥분제	healthywoman	운동하는여자
28	vloga	비로그라	healthyman	운동하는남자
29	clalls	시알리스	tripstagram	여행스타그램
30	sexdrug	섹스파는곳	moviestagram	영화스타그램
31	smallloan	소액대출	coffeestagram	커피스타그램
32	loanprod	대출상품	termexam	시험기간
33	hotmoney	급전	homework	과제
34	pawmloan	내구제	weather	날씨
35	conditional.meeting	조건부만남	picnic	소풍
36	sex.tag	섹태그	walk	산책
37	sexgode	섹고드	recent	근황
38	sexcode	섹코드	festival	축제
39	sadaripic	사다리픽	TRUE	진심
40	named.ladder	네임드 사다리	family	가족사진
41	named	네임드	weekend	주말
42	casino	카지노	wiseSaying	행언
43	safe.park	안전공원	textstagram	글스타그램
44	pickster	픽스터	student	학생
45	betting	배팅	library	도서관
46	private.gambling	사설놀이터	café	카페

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

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 (노동의 시간..)	20,530 records ➔ 17,705 records (노동의 시간..)

3. Data preprocessing

- 2차 전처리로 KoNLPQ를 사용하여 시행함
- Supervised term weighting method중 DF기준으로 Prob_W와 Entropy기반의 Information Gain을 사용함
- KoNLPQ는 TagSet중 Email, URL을 지원하기 때문에 3가지에 대한 Meta 변수를 추출함



POS tagging
'KoNLPQ'
(Twitter POS Tagger)

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

패키지	결과
Kkma (Python)	[(*, 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), (ㄴ니다, EFN), (., SF)]
KoNLPQ (R)	*(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)

Supervised term weighting method
based on contingency table

Feature \ Category	C_j	\bar{C}_j
f_i	$f_i C_j$ (1)	$f_i \bar{C}_j$ (2)
\bar{f}_i	$\bar{f}_i C_j$ (3)	$\bar{f}_i \bar{C}_j$ (4)

- (1) $f_i C_j \rightarrow C_j$ 클래스에서 f_i term이 있는 document 수
 (2) $f_i \bar{C}_j \rightarrow C_j$ 가 아닌 클래스에서 f_i term이 있는 document 수
 (3) $\bar{f}_i C_j \rightarrow C_j$ 클래스에서 f_i term이 없는 document 수
 (4) $\bar{f}_i \bar{C}_j \rightarrow C_j$ 가 아닌 클래스에서 f_i term이 없는 document 수

따라서 (1)+(2)+(3)+(4) = 전체 document 수

사용한 형태소 모양

Adjective
Adverb
Alpha
Foreign
Koreanparticle
NA
Noun
Verb

Weight method

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

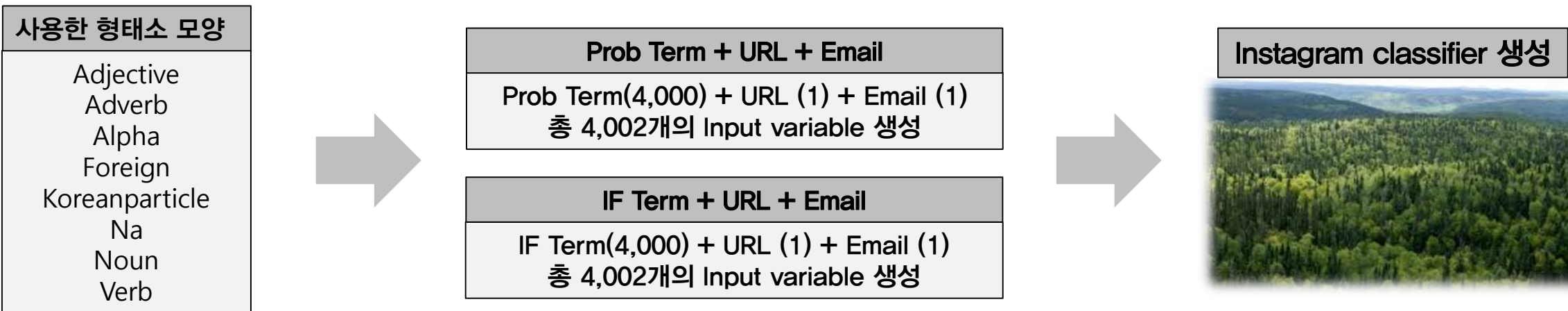
$$W(f) = tf * \log \left(1 + \frac{f_c f_{\bar{c}}}{f_{\bar{c}} \bar{f}_c} \right)$$

TF-PROB equation

3. Data preprocessing

- Instagram에서 생성된 전체 Term은 144,187개임, 하지만 굉장히 sparse함
- 모든 Term를 사용 하는 것은 PC 환경상 비효율적이며, 성능도 좋지 않음
- Prob weight, Information Gain기준 상위 4,000개만 사용함

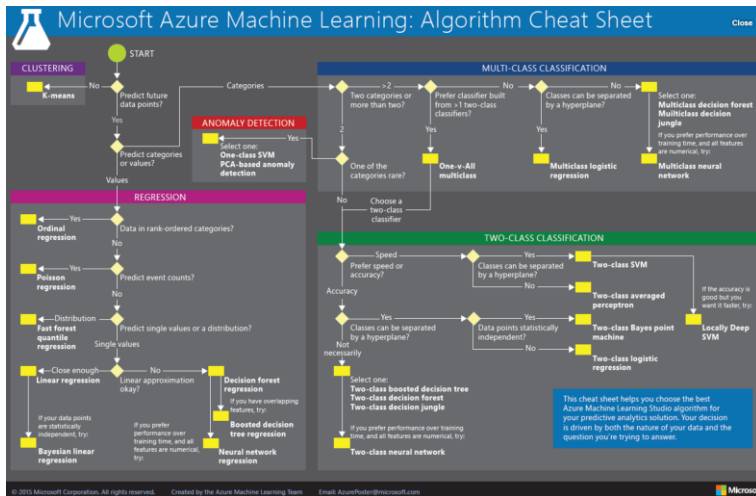
Term	Spam_TF	Ham_TF	Total_TF	Spam_DF	Ham_DF	Total_DF	Num_of_Doc	Num_of_Spam_Doc	Num_of_Ham_Doc	CM_1_1	CM_1_2	CM_2_1	CM_2_2	Chi_sq_W	Prob_W	RF_W	VRF_W
메이저zqqnoun	1813	0	1813	1787	0	1787	203002	20665	182337	0	1787	182337	18878	0.0000534026943724	5.13778839825791	0.739414358085732	2.35695251646158
스포츠토zqqnoun	2724	2	2726	2508	2	2510	203002	20665	182337	2	2508	182335	18157	0.0003176858455420	4.75841648707327	0.759953134085185	1.97922696756528
사설zqqnoun	3516	6	3522	2726	5	2731	203002	20665	182337	5	2726	182332	17939	0.0011412944718870	4.24973790011247	0.766400750548987	1.8838307527672
메이저zqqnoun	1713	2	1715	1507	2	1509	203002	20665	182337	2	1507	182335	19158	0.0005853681429181	3.70289983355734	0.731747386222628	2.54198830952809
강남풀싸롱zqqnoun	1430	0	1430	863	0	863	203002	20665	182337	0	863	182337	19802	0.0001211131133584	3.65573899162882	0.714727517513031	3.13201595155282
네임드zqqnoun	9328	45	9373	3919	29	3948	203002	20665	182337	29	3919	182308	16746	0.0173044818178061	3.45269430704749	0.803825828174372	1.45212748394997
사다리zqqnoun	9095	33	9128	3333	20	3353	203002	20665	182337	20	3333	182317	17332	0.0106629033522763	3.45118774670497	0.784973960767212	1.64843941463057
섹스zqqnoun	11558	95	11653	6115	87	6202	203002	20665	182337	87	6115	182250	14550	0.0723742825739593	3.40823507465256	0.883897595760445	0.866751431711326
풀싸롱zqqnoun	3080	2	3082	1019	1	1020	203002	20665	182337	1	1019	182336	19646	0.0004037659983303	3.31336111117449	0.718774148262042	2.95812202760785
바카라zqqnoun	16375	324	16699	4262	50	4312	203002	20665	182337	50	4262	182287	16403	0.0440124164318217	3.1233576540734	0.81530978526035	1.34755206974674
카지노zqqnoun	23078	486	23564	4846	72	4918	203002	20665	182337	72	4846	182265	15819	0.0737582932712828	3.06072755039345	0.835681070423778	1.18291500547627
사이트zqqnoun	23071	449	23520	5856	144	6000	203002	20665	182337	144	5856	182193	14809	0.210191107207184	2.83171246414486	0.873581970908031	0.927665101172062
야동zqqnoun	5219	76	5295	3584	48	3632	203002	20665	182337	48	3584	182289	17081	0.052125957834961	2.79452004528266	0.792933707753985	1.56126680347237
일탈녀*zqqnoun	3668	45	3713	3263	39	3302	203002	20665	182337	39	3263	182298	17402	0.0395916125929114	2.79143475205842	0.782783965242369	1.67368916805018
토토zqqnoun	10682	129	10811	4686	96	4782	203002	20665	182337	96	4686	182241	15979	0.136562065511234	2.71947367863914	0.829993629255384	1.22654522139756
배팅*zqqnoun	2711	5	2716	1275	5	1280	203002	20665	182337	5	1275	182332	19390	0.0028244232668767	2.70766736148212	0.72551936297856	2.72107885597788



4. Modeling

– Text classifier

- Classification 문제에서 가장 범용적으로 성능이 우수하다고 알려진 ‘RandomForest’ 를 사용하기로 함
- 순서가 상관없는 Implicit ensemble기반이기 때문에 parallel computing이 가능함 (parallel + randomForest) [R]
- Supervised Term weight의 전처리 필요 없음
- Distance를 Similarity로 고쳐줄 필요도 없음



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

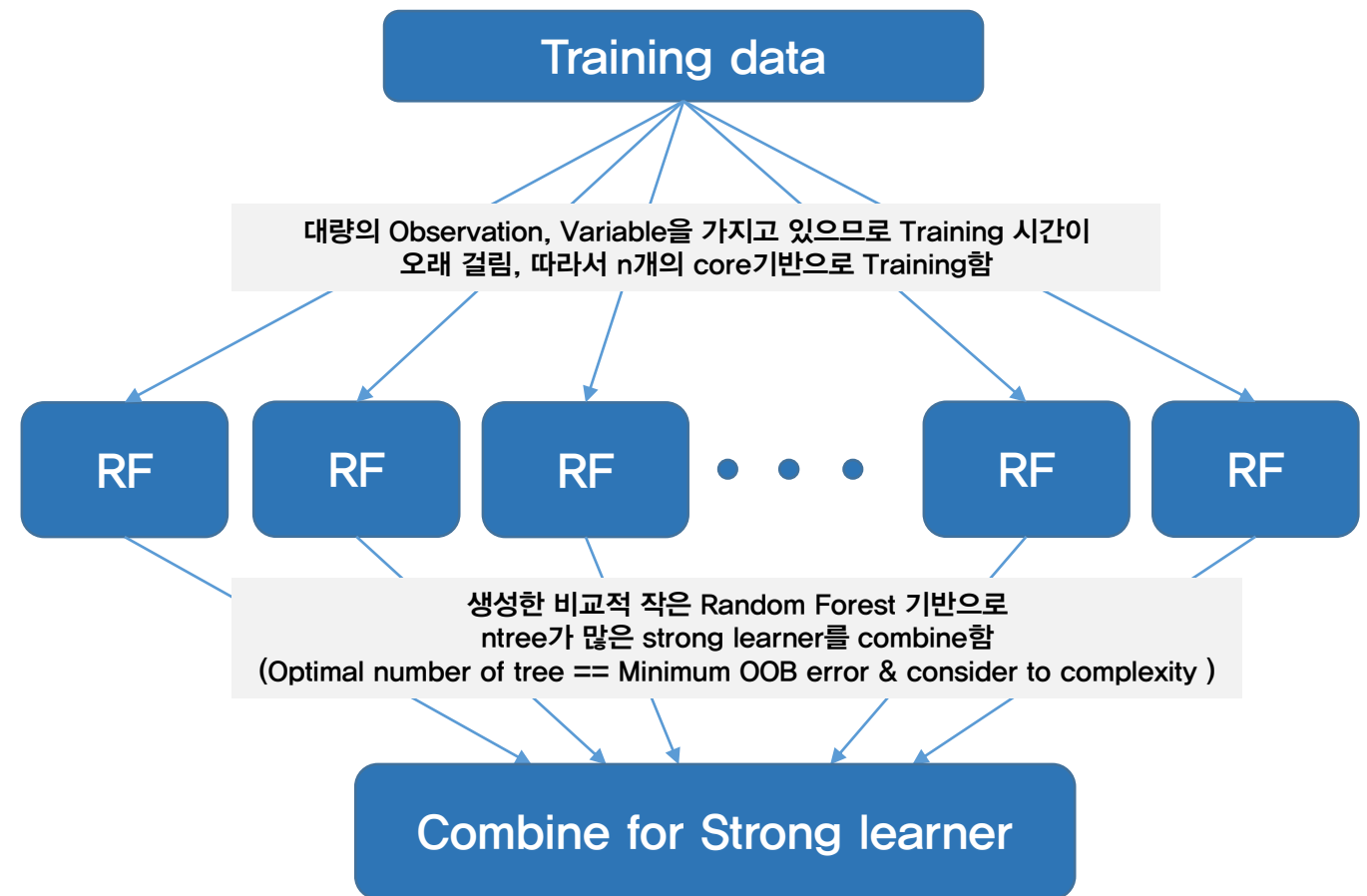
Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, Dinani Amorim, 15(Oct):3133–3181, 2014.

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]

<http://jmlr.csail.mit.edu/papers/v15/delgado14a.html>



4. Modeling

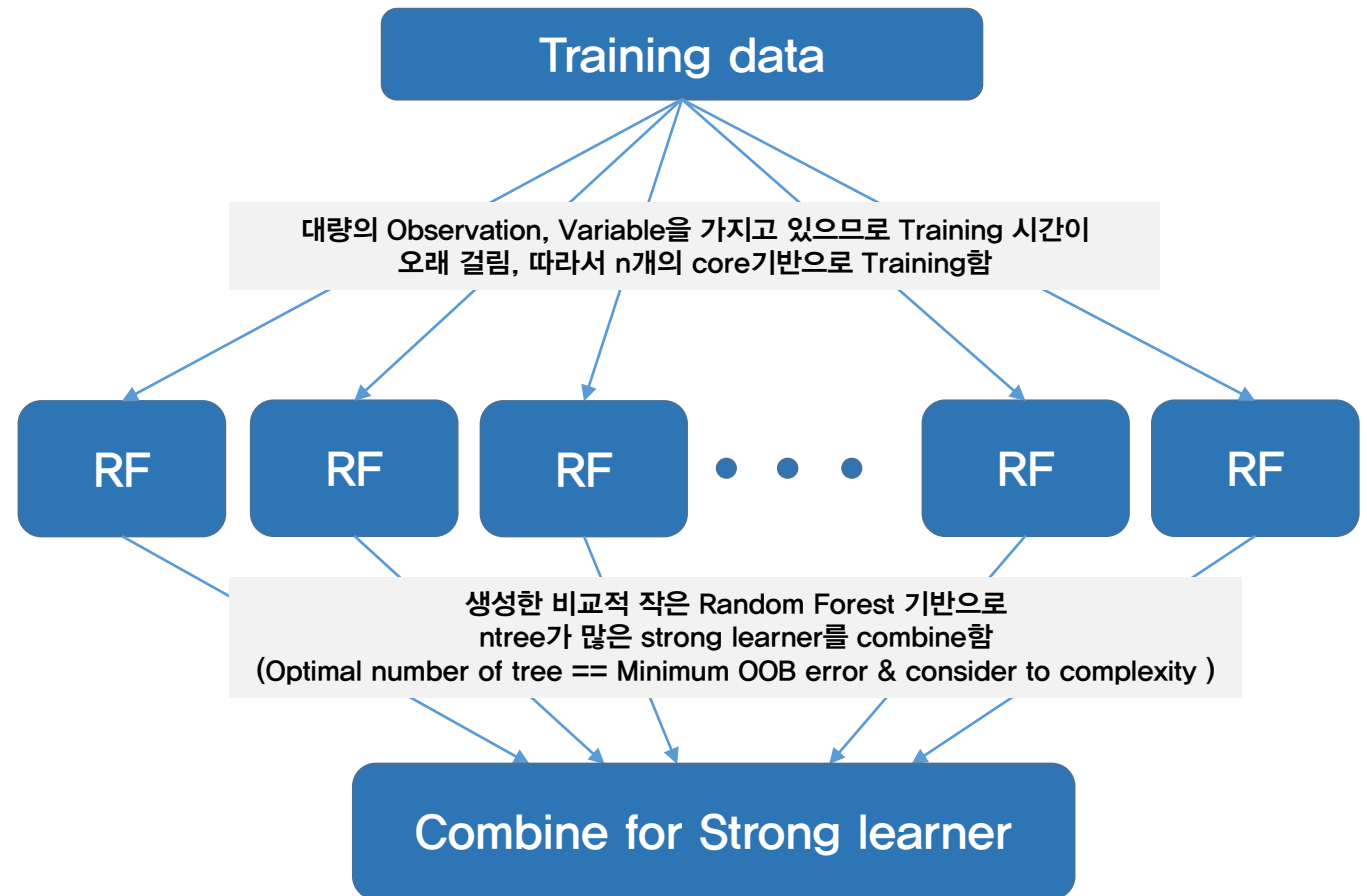
– Text classifier

- RandomForest는 unbalanced data에 민감한 특성과 그대로 사용할 경우 PC에서는 굉장한 메모리를 필요로 함
- Down sampling을 하여 모델링을 하였음 (HAM 25,000개, SPAM 17,705개)
- 4,002개의 벡터는 HAM경우는 SPAM경우 보다 비교적 훨씬 sparse한 구조임
- 변수의 수는 보통 $\sqrt{4,002}$, 약 63개를 사용하지만, sparse한 구조이기 때문에 200개를 사용함



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

<http://jmlr.csail.mit.edu/papers/v15/delgado14a.html>



4. Modeling

– Text classifier

- Information Gain기반이나 Prob_W 기반이나 큰 차이 없이 OOB Error기준 약 2%정도의 error를 산출함
- SPAM Term을 기준으로 가져온 Prob_W방식을 채택함 (One class)
- 부가적으로 Instagram에서 SPAM URL을 생성할 수 있음



Instagram

Training

1

Construct Random forest



Type of random forest: classification
Number of trees: 65
No. of variables tried at each split: 200

OOB estimate of error rate: 1.89%

Confusion matrix:

	0	1	class.error
0	24434	566	0.02264000
1	23917	466	0.01349901

Type of random forest: classification
Number of trees: 65
No. of variables tried at each split: 200

OOB estimate of error rate: 1.72%

Confusion matrix:

	0	1	class.error
0	24494	506	0.0202400
1	22917	476	0.0129342

Prob기반 Term사용

산출가능 한 2가지

1. Classifier
2. Spam URL 집합

facebook

4. Modeling

– Text classifier

- 지금까지, Instagram 기반으로 RandomForest를 생성하였음
- 이 모델을 기반으로 Facebook을 예측하며, 확률값을 도출함

이제 1 번은 완성되었고

2 3 번을 해야함



4. Modeling

– Text classifier

- Instagram으로 만든 RandomForest로 동일하게 4,002개의 변수로 분류를 해봄
- Facebook 데이터는 총 1,795,067개임
- 0.1 이하로 1,749,428개이며, 0.1초과로 29,397개임
- 0.1 이하는 전부 HAM으로 간주하였음

2180389	105563005	유성호 이거봐 나도있지도않는데 오빠는화석임	유성호 이;유성호 이 유성호ZQ(유성호 이 유성호ZQ(0	0	0	0.061538
2182532	105728700	심소령 권유진 촬영은 정혜경 가자 솔로부대여	심소령 권;심 소령 ;심ZQQNo 심 소령 ;심ZQQNo	0	0	0	0.061538
2186741	106140675	백정금 ..?대박 이래서서울로갔어야했어ㅋㅋ	백정금 ;백정 금 ;백정ZQQ(백정 금 ;백정ZQQ(0	0	0	0.061538
2192910	106454573	김규리 배운진 조유빈 족보가뭐죠? 먹는건감?	김규리 배;김규리 배 김규리ZQ(김규리 배 김규리ZQ(0	0	0	0.061538
2197923	106819015	최효정 사지영 시급이 3만원????????	최효정 사;최효정* 사최효정*ZQ최효정* 사최효정*ZQ	0	0	0	0.061538
2205848	106937925	서윤해 김민경 학식이랑 깃사가격 올랐을때—	서윤해 김;서윤 해 ;서윤ZQQ(서윤 해 ;서윤ZQQ(0	0	0	0.061538
2218129	107965558	Mijin Sim 이중에 가격왕을... ?!	Mijin Sim Mijin Sim MijinZQQ Mijin Sim MijinZQQ	0	0	0	0.061538
2219309	108275811	이태연 뭐야 초코파이랑 투게더 먹어보고 싶음	이태연 뭐;이태연 뭐 이태연ZQ(이태연 뭐 이태연ZQ(0	0	0	0.061538
2224601	108895873	김진호 손은 부러졌어도 Y존케어는 잘했나보군	김진호 손;김진호 손 김진호ZQ(김진호 손 김진호ZQ(0	0	0	0.061538
2226852	109274895	백재민 오 역대급이다 사진 잘 모아둬ㅋㅋ	백재민 오 백 재민 ;백ZQQNo 백 재민 ;백ZQQNo	0	0	0	0.061538
2228894	109862644	박지민 요즘 그렇게 공부를 열심히 한다며..	박지민 요;박지민 요 박지민ZQ(박지민 요 박지민ZQ(0	0	0	0.061538
2248835	432013253	박소영 개귀엽ㅋㅋㅋㅋㅋㅋ정녕 27맞는가	박소영 개;박소영 개 박소영ZQ(박소영 개 박소영ZQ(0	0	0	0.061538
2249584	432013253	이예지 내사랑♡ 우리타짜보러가자탑보러ㅋㅋ	이예지 내;이예지 내 이예지ZQ(이예지 내 이예지ZQ(0	0	0	0.061538
2250009	432168866	이혜랑 힐 나저거 사출사람?? 업겠지...?	이혜랑 힐;이혜 랑 ;이혜ZQQ(이혜 랑 ;이혜ZQQ(0	0	0	0.061538
2254808	564438293	박미정 조민정 강연지 김은지 찍으러가자 ㄱㄱ	박미정 조;박 미정 ;박ZQQNo 박 미정 ;박ZQQNo	0	0	0	0.061538
2254883	564438293	원혜린 김원 혜린 ㄱㄱㄱㄱㄱㄱㄱㄱㄱㄱㄱㄱㄱ	원혜린 김;원 혜린 ;원ZQQNo 원 혜린 ;원ZQQNo	0	0	0	0.061538
2255619	564438293	이휴정 진심보고싶는데 오빠랑보지말고 나랑보자	이휴정 진;이 휴정 ;이ZQQDe 이 휴정 ;이ZQQDe	0	0	0	0.061538
2262449	604176773	황주현 우린 노래방이나 가자ㅋㅋㅋㅋㅋㅋ	황주현 우;황주현 우 황주현ZQ(황주현 우 황주현ZQ(0	0	0	0.061538
2262710	604176773	Kyung Jae Yoo 자기야 눈사람 ㄱㄱ?	Kyung Jae Kyung Jae KyungZQ(Kyung Jae KyungZQ(0	0	0	0.061538

SPAM Term기준으로 만든 Instagram RandomForest 적용 결과 중 0.1 이하는
Facebook에서는 HAM으로 인식할 만큼의 결과만 산출이 되었음

➡ Prob_W 로 만들었기 때문에 Spam Term이 없는 0 Vector인 경우가 비교적 많음

- Text classifier

- #5년차, 메이저 놀이터 문의: boss757
- ★모든 검증사이트 검증완료★
 - ★다른 놀이터의 거짓된 정보와 상술에
 - ▷더이상 피해보지 마세요!!
 - ▷먹튀걱정없는 메이저 놀이터
 - ▷5년차 믿음직한 놀이터!!!
 - ▷매주 다양한 이벤트 진행중
 - ▷메세지로 이벤트 실시간 전송
 - ▷페이소북에 주소, 코드는 놀이터
 - ▷다 먹튀사이드 입니다. 조심하세요!!
 - ▷신규이벤트 3중전시 + 1실미디게임
 - ▷크로스배팅은 기본, 한디/언오버
 - ▷실시간 LIVE 배팅(사다리, 달팽이, 파워볼)
 - ▷사다리 조합배팅 가능
 - ▷사다리연승 이벤트 진행 (5승, 7승, 10승)
 - \$ 빠른 중/환전 & 24시간 고객센터운영
 - ▷vp-sk.com // 코드 toto647
 - ▷문의:boss757

아여진구오빠이제..성인이네
미안나먼저시집갈께김주이강유진

@성인제 김윤 저셋기 존니 태그하고싶는데 비활성화됨 ㅋ
Kyung Jun Roh 도박 마약ㄸㅋㅋㅋㅋㅋ짱웃가
송성민 김현준 내가 강화를하는이유 섹스2번차리임 ㅇㅇ
조연빈 민주익언덕에 우리가 간곳은 민주익 성인돗?ㅋㅋ

그보다 섹스에환장한암컷이 엄청거슬리네요.
암컷???

→ 어떻게 하면.. 무엇이 문제일까..?

4. Modeling

– Text classifier

- 분류 0.1이상으로 나타낸 원래 문자열의 결과임
- Facebook에서는 성능이 그렇게 좋지 않음 → 하지만 이것을 극복하였고 그 결과는 아래 표와 같음

#5년차, 메이저 놀이터 문의: boss757
 ★모든 검증사이트 검증완료★
 ★다른 놀이터의 거짓된 정보와 상술에
 >더이상 피해보지 마세요!!
 >먹튀걱정없는 메이저 놀이터
 >5년차 믿음직한 놀이터!!
 >매주 다양한 이벤트 진행중
 >메세지로 이벤트 실시간 전송
 >페이스북에 주소, 코드없는 놀이터
 다 먹튀사이트 입니다. 조심하세요!!
 >신규이벤트 3중전시 + 1실머니지급
 >크로스배팅은 기본, 핸디/언오버
 >실시간 LIVE 배팅(사다리, 달팽이, 파워볼)
 >사다리 조합배팅 가능
 >사다리연승 이벤트 진행 (5승, 7승, 10승)
 \$ 빠른 중/환전 & 24시간 고객센터운영
 ▶vp-sk.콜 // 코드 toto647
 ▶문의:boss757

아여진구오빠이제..성인이네
 미안나먼저시집갈게김주이강유진

@성인제 김윤 저셋기 존니 태그하고싶은데 비활성화잼 ㅋ
 Kyung Jun Roh 도박 마약ㅋㅋㅋㅋㅋㅋ짱웃가
 송성민 김현준 내가 강화를하는이유 섹스2번짜리임 ㅇㅇ
 조연빈 문주익언덕에 우리가 간곳은 문주익 성인듯?ㅋㅋ

그보다 섹스에환장한암컷이 엄청거슬리네요.
 암컷???

분류의 좋은 예시 ☺

분류의 좋지 않은..예 :<

→ 어떻게 하면.. 무엇이 문제일까..?

답은 Facebook 특성에 맞는 Meta 변수!

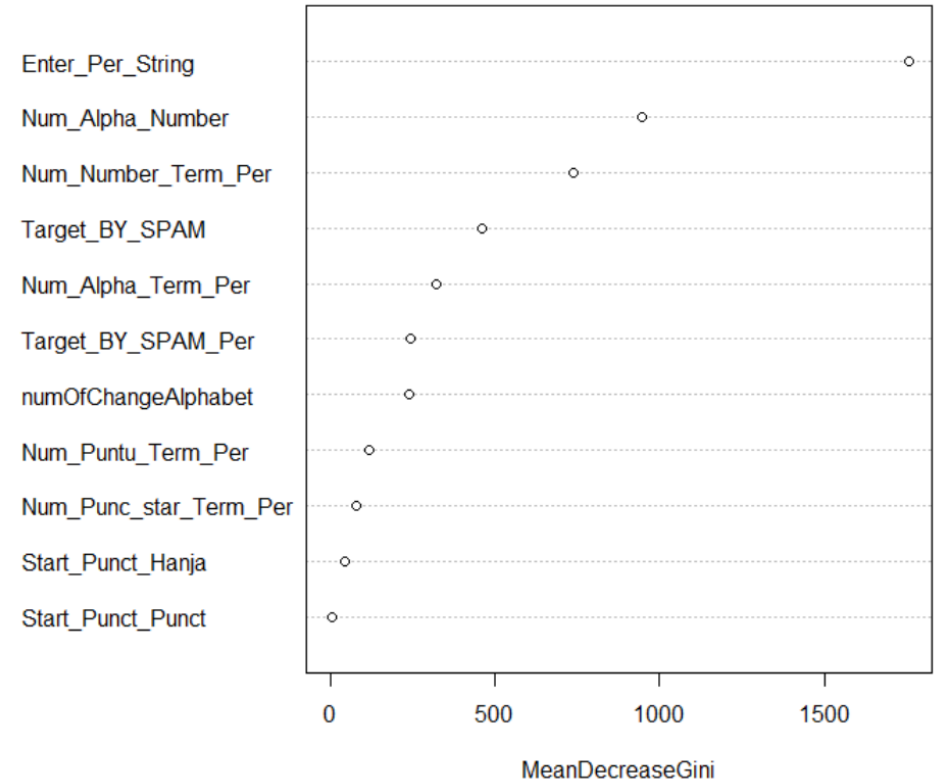
Accuracy		F1 measure		BCR		Specificity		Sensitivity		Recall	
Instagram RF	Instagram + Facebook RF	Instagram RF	Instagram + Facebook RF	Instagram RF	Instagram + Facebook RF	Instagram RF	Instagram + Facebook RF	Instagram RF	Instagram + Facebook RF	Instagram RF	Instagram + Facebook RF
0.94279	0.99344	0.678113	0.965242	0.83583	0.983093	0.962499	0.995817	0.726172	0.970576	0.637072	0.960174
0.004506	0.001772	0.023353	0.009077	0.017935	0.006457	0.003679	0.001572	0.031444	0.012816	0.030432	0.0143

4. Modeling

– Text classifier

- Facebook을 위한 총 10가지의 파생 변수를 생성함
- 해당 변수를 와 Instagram에서 만든 Learner의 P값을 더하여 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

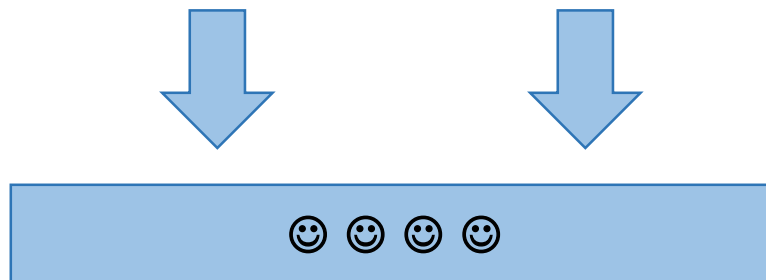


4. Modeling

– 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

Accuracy		F1 measure		BCR		Specificity		Sensitivity		Recall	
Instagram RF	Instagram + Facebook RF	Instagram RF	Instagram + Facebook RF	Instagram RF	Instagram + Facebook RF	Instagram RF	Instagram + Facebook RF	Instagram RF	Instagram + Facebook RF	Instagram RF	Instagram + Facebook RF
0.94279	0.99344	0.678113	0.965242	0.83583	0.983093	0.962499	0.995817	0.726172	0.970576	0.637072	0.960174
0.004506	0.001772	0.023353	0.009077	0.017935	0.006457	0.003679	0.001572	0.031444	0.012816	0.030432	0.0143



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

4. Modeling

– Text classifier

- 반복실험으로 멈추지 않았으며 최근 데이터로 실제 적용해 보고 싶었음
- 실제 남자들의 동영상 페이지에서 다시 크롤링을 하여 이를 실험해봄 (2016.04.01 ~ 2016.05.31)
- 다른 페이지에 비하여 스팸 비중이 높아보였기 때문에 선정함



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



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



4. Modeling

– Text classifier

- Instagram RandomForest로 먼저 0.1이하인 경우 HAM으로 예측하였고 이를 평가해 보았음
- 2,176개 중 2,115개가 여기에 해당함
- 전부 HAM인것 을 확인 하였음

이동환 니가 만나고싶어하는여자 맥스찍었다
이봉식 봉식아 왜 이런거에만 따봉눌러?
정준호 너가 좋아할거같아서 ... 취향저격ㅋㅋ?
박종현 이것도사랑ㅋㅋㅋㅋㅋㅋ
김성훈 윤영훈 동생이 형들을 위한 마음입니다 ㅋㅋㅋㅋ
차성호 애 이럴줄알았어
이중도 수고
홍정표 꿀
저거 다 해당되면 그냥 개 폐인 새긴 거자날ㅋㅋㅋㅋ
김민철 어딜봐서 평등이죠 이런게?ㅋㅋ
김진서1번부터 전국 50명제한
조수빈 우리 이 피씨방 봤잖아 ㅋㅋ
김지수 아마 나 못함~
이거 면접보려면 나이 한 40대는 되겠다
Hwa Woong Kim 불가능 ㅇㅈ? 우리 스펙 다더해도 불가능
천준오 추동수 이동재 8차시험 통과해본다 내가
뭐지..시발 서든이 찰쉬워보인다?
황대우 1차부터 탈락이그..☆
서재경 구단까지 꺾값인데?
1단계가 제일 어렵네
2,3해당이네
더러운ㅇ세상.
메이플토리
더러운 세상 씨바
블소 진짜만렙 20성으로 치면 블소는 불가능인데 이미
메이플 가능성 있네 ㅋㅋ
1차에서 이미 떨어진 사람 개많음

알바할빠에야 가지고 있는 계정 파는게 더 돈되겠네
김광훈 알바도 여자만 뽑으니 소소주우
김진현 알바줄구해조ㅋㅋㅋ괜히 나왔나바... ㅌㅌ
성지석 지석아 가능?
양준민가능?
봄침번이 불을? 아무리 신막사 동기생활관이라도 다동기끼리 생활관을
아침에 일어나면 불키고 바로 티비를킨뒤 유비하나 틀고 커튼 치고 창문
손세욱 뽀소리임 늦잠잔거?
저 정도면 알바가 아니고 프로게이머를 해야할듯 ㅋㅋ
술까 피방은 이쁜 여알바있으면 거기간다. 다 거기서 거기기 때문에
저거 다 되면 피방에서 왜 알바하나 강 차리고말지 ㅋㅋㅋㅋ
저게다있으면알바를왜함
장난하나 아무리 웃으라고 한거지만
여자만 뽑는다하던가 눈살 찌푸러지게 뭘 장난이고? 뭐 시급 5만원하
나???
이다경너내알바도 면접이래?
리니지 레벨70이라는거부터 알바를 할필요가없다는거아닌가....
그 알바 시급 얼마? 역대연봉???
지대찬 대찬아 저기 알바뽑는데
ㅋㅋ 아재들 군대이야기에 미치시네요.. 요즘 그냥 일찍일어나사람이
저런...이동병친구 꿀잠잔네
김지훈 이걸 보면 댕 긴머리냐?
ㅋㅋㅋㅋ시발 그냥 댕 피시방 알바면접보겠다
피시방알바에서 신재검사는 왜하는덱ㅋㅋㅋㅋ
이승미 알바하러갈래?
남자가 저거 다 만족하면 알바를 멀리하고 아프리카를 하는게 낫습니다

최명식 ㅋㅋㅋㅋ
안다옥 ㅋㅋㅋㅋ
양우합 ㅋㅋㅋㅋ
김효겸 ㅋㅋㅋㅋㅋㅋ
현인호 황진홍 ㅋㅋㅋㅋㅋㅋ
다음생엔 여자로 태어나는걸로ㅋㅋ
진짜 잇어라 ㅋㅋㅋ 몇개는 되는데
시급 얼마 줄라고 ㅋㅋ
정복자레벨2천 ㅋㅋㅋㅋㅋㅋ
이유빈 ㅋㅋㅋㅋ
손민우 ㅋㅋㅋㅋ
박준형 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
ㅇㅈ ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
박종현김찬영ㅋㅋㅋㅋ
이동병이 항상 먼저일어나서 불을 켜야되는데 눈떴을때 누군가 고참에
이기재 이거 늦게 일어난거예요??ㅋㅋ
이태희 ㅋㅋ 너가 먼저 꼭 켜라잉
김한 너의 미래구나 ㅋㅋ
배영광 개망 ㅋㅋ
장현진 넌 안키냐 ㅋㅋ
최선희 알아? ㅋㅋ
김정아 뭔지알아?ㅋㅋ
김하늬이해가??ㅋㅋ
기상이 올리면 재빠르게 일어나서 불키고 커튼치고 침구류 개고 환복한

조현영 정재윤 김준하
이상민 정충만
박민창 김준영
ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
Young Jae Cho JaeYeon Chun Jason Uhm Justin Lee
이세라
이가현
남승화
박준서
황빛나
Lewis Kim
이정은
슬기
정성희
김호진
장예슬
Hyunjung Kim
신휘정
정상훈
임한
이명훈
차민혁
박예슬
백예지
김주희
최정인

말끔하게 HAM으로 분류가 된 것 을 확인 할 수 있음

- Text classifier

- | | | | |
|---------|------------|--|--|
| results | FullString | 이벤트공지 *****

* 최소 배팅 3천원 부터 가능

* 핸드사다리 이벤트 5연승 10연승
(5연승 연승금액 30% 지급,
10연승 연승금액포함300% 지급

* 삼만원입금시 1만더 지급

* 매일 첫 충전 10 % 지급

1 * 풍포인트 4900포인트 지급
(5000 포인트부터 캐시 가능)

* 핸드사다리 분석기 무료 지급
(조건없이 필요하신분 무료지급)

* 매일 첫 이벤트 운영중인 카카오톡방 운영
3만 중전시 3+1 + 4900 + 첫인금보너스 =

* 47900원으로 배팅고교성 * | 이벤드공지

* 최소 배팅 3천원 부터 가능

이벤트공지

* 최소 배팅 3천원 부터 가능

이벤트공지

* 최소 배팅 3천원 부터 가능

* 처음충전 4900 원 +포인트 지급
3년선호준부 주사코메리준등 단인출이며 Queen!!
추가회원님들을 환영하며
다양한 혜택 진행중!
시간유저 특허 천인금보너스15%

3+1첫회화하지그
저항 만일 하실 오빠들 카톡 nbnb58 친추용^^
★ 레미스를 넷플 알바보신분 전로 역귀됩니다

페이스북에서 홍보중인 놀이터는 꼭 검증후에
이용하세요
♥ 만티배드는 사옹주이 어찌이 건중 피해사례를
♥ soa101.컴 (컴=com) ♥ 진짜 거짓말안하고 다른 사이트와 비교도 안되는 곳입니다! 늦기전에 이쁜 파트
♥ soa101.컴 (컴=com) ♥ 진짜 거짓말안하고 다른 사이트와 비교도 안되는 곳입니다! 늦기전에 이쁜 파트
♥ soa101.컴 (컴=com) ♥ 진짜 거짓말안하고 다른 사이트와 비교도 안되는 곳입니다! 늦기전에 이쁜 파트
♥ soa101.컴 (컴=com) ♥ 진짜 거짓말안하고 다른 사이트와 비교도 안되는 곳입니다! 늦기전에 이쁜 파트
♥ soa101.컴 (컴=com) ♥ 진짜 거짓말안하고 다른 사이트와 비교도 안되는 곳입니다! 늦기전에 이쁜 파트
♥ soa101.컴 (컴=com) ♥ 진짜 거짓말안하고 다른 사이트와 비교도 안되는 곳입니다! 늦기전에 이쁜 파트
각종 영업도우구 지원
각종 영업도우구 지원
(네이트 사다리,UFC,스타2등등)
***** 이벤트공지 *****

* 최소 배팅 3천원 부터 가능

* 캐쉬주는 이벤트 5연승 10연승
(5연승 연승배팅금액 30% 지급,
매니저 모집(캐피)
각종 소비재 지원
각종 도구지원
(실시간 사다리,UFC,스타2등등)
매니저 보신(영업번)
각종 SNS 아이디 지원
각종 영업도우구 지원
(실시간 사다리,UFC,스타2등등)
외곽 7억 적자 처리 |
|---------|------------|--|--|

* 캐쉬주는 이벤트 5연승 10연승
(5연승 배팅하신돈의 30% 지급,
(10연승 배팅하신돈의포함300% 지급

길이 가변적인 SPAM도 모두 예측한 것을 볼 수 있음

4. Modeling

– Text classifier

- Instagram RandomForest로 0.1초과의 경우 기존의 Facebook Meta 변수를 사용한 RandomForest를 이용함
- 2,176개 중 61개가 여기에 해당함
- 가변적인 길이에도 분류 결과가 우수하게 나타나는 것을 알 수 있음 (☺ ☺ ☺)
- 61개에서 37개를 햄으로 27개를 스팸으로 예측함 27개는 확실히 정보가 있는 스팸이며 예측한 데이터가 모두 스팸임을 확인

ㅋㅋ 남자이거 풀고 들어가면
거의 삼성 혹은 그 이상 가는 스펙일듯. 게임업계에서ㅋㅋ
정래범 김민석 신동진 정한솔 사법고시보다 뽕센 pc방 알바 4급공무원말고 4급알바 ㄱ ㄱ
아이디 파는게 돈 더벌리겠다
저정도 스펙이면 게임회사에서 스카우트해가죠 서로 대려가겠다고 프로그래머도 가능하겠네요..를 챌린저면..ㄷ
솔직히 여자는 면접없어도 가능하다
문열고들어오는순간 바로시간및시급상당임 서유진
저스펙에 피방알바를왜함 최향진 ㅇㅈ? 으인정
그냥그게임 계정을팔고말지ㅅㅂ
남자 통과할수 있다면 그사람은 게임으로 돈벌어먹을수 있겟ㅋㅋㅋㅋㅋㅋ
저정도했으면 강 아이디를 팔았겠다
윤태경 전요환 가능??ㅋㅋㅋㅋ
이정도 스펙이면 게임회사를 가겠다
4차 던파 풀에픽 현금 오백만원가량질러야가능함
아이디만팔아도 답나옴
신체검사까지가면 게임중독으로 찌르겠지? 사장 개객기
미친섹파구하냐??피시방면접조까이하네
한국회 이젠 알바도어려워
김대창 가능?
중고폰치고가매입 <https://m.facebook.com/TAESUNGIT/>
언제부터 불침번이 붙을켰지?ㅋㅋ
막내가 불키고 커튼열고 m.net틀지않나?
유환년 이장면을 보았지ㅋㅋ

하호준 우린 이런 적 있었던 게 행정실 방송기계가 고장나서 기상나팔을 못 틀었었거든 ㅋㅋ
Kitaek Kim 늦잠잤지말입니다?? 졸라빠져가주고
좀더 각도가 밀인듯한데...막내는 불바로 밀에자지
와 생각만으로도 후덜덜 군생활꼬이는 소리가 들린zz
이유영 불키고 커튼 걷고 티비 키고 환복하고 걸레 빨로 가던 그 곳
권성욱 ㅅㅂ자는데 누가불켜놔노
가로본능으로 잤네ㅋㅋㅋㅋ ㅋㅋㅋㅋㄱㄱ
It의모든것 <https://m.facebook.com/TAESUNGIT/>
www.facebook.com/haejuoentertainment 좋아요 부탁드립니다 :)
어린이는 출입금지!!! 성인만오세요 ^ ^!!!
한진혁 신상엽 김병준 잘가 혁동..
신용근
예쁘다 내가 지켜줄게요언니
물들어올때 노젓ㅅ는거냐
어린이는 출입금지!!! 성인만오세요 ^ ^!!!
많이 예쁘시고, 섹시하세요.
양정원 요새모하고다니는거녀
김준제 너가조아하는여자다

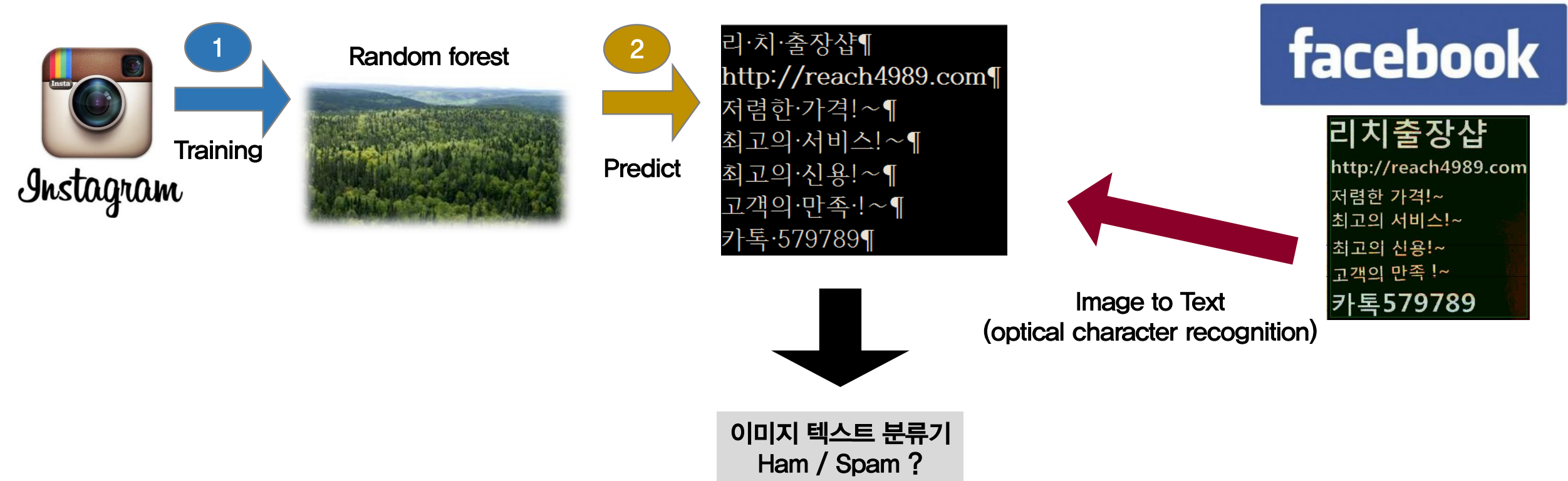
중고폰광고 이외에는 어떤 URL이나 정보를 담은 SPAM성 글은 없음

➔ 결국 2176개중 1개를 틀리고 잘 분류함

4. Modeling

– Image classifier

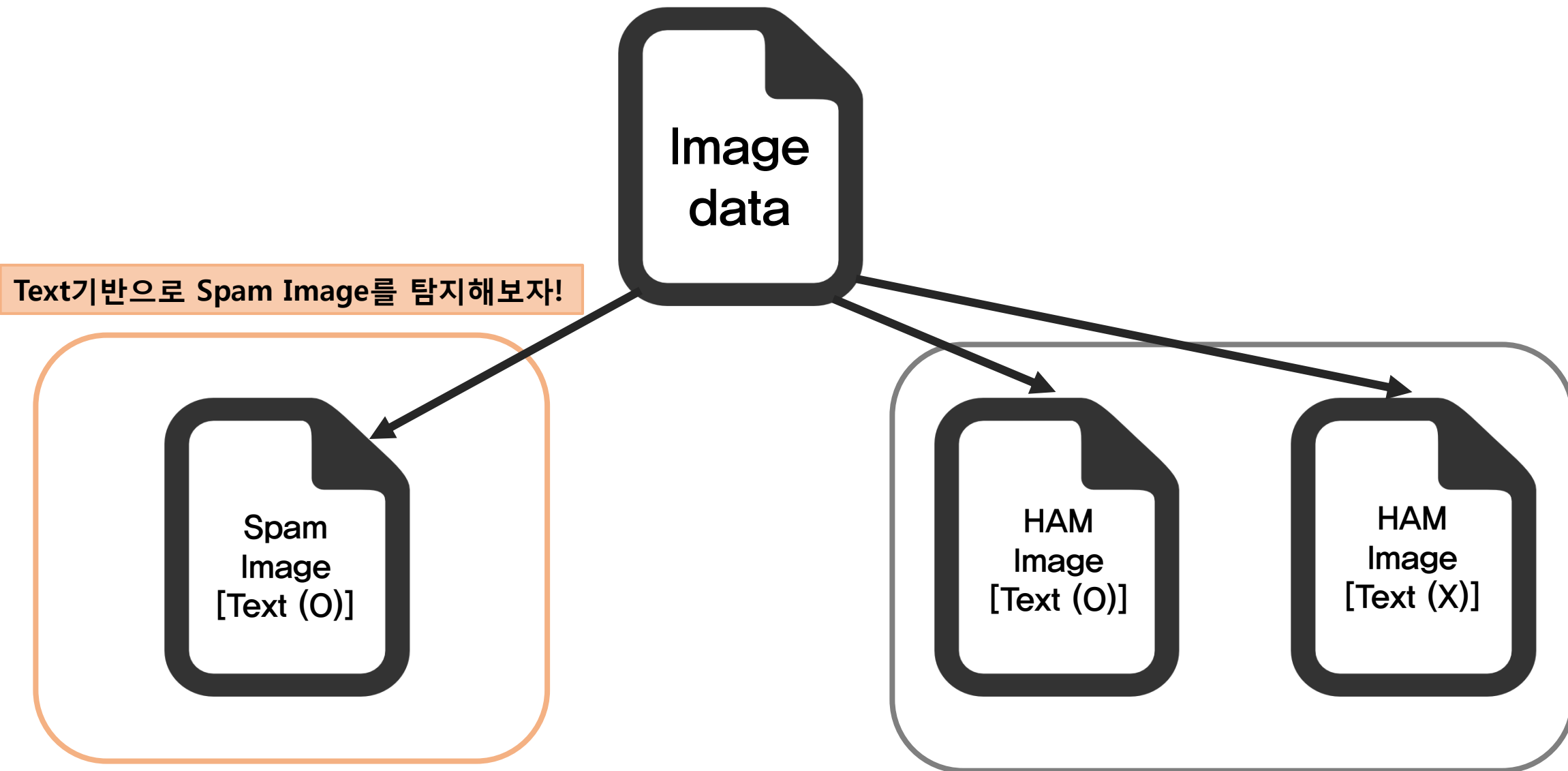
- OCR을 통해 이미지에 있는 텍스트를 추출함
- PROB_W를 통해 생성한 Random forest를 사용하여 추출한 텍스트를 통해 Spam/Ham을 classification함



4. Modeling

– Image classifier

- 연구 모형에서 Spam image의 정의는 Spam 정보를 포함한 텍스트가 포함된 이미지를 의미함



4. Modeling

– Image classifier

- Text가 없는 Ham Image들은 텍스트가 추출되지 않음



<내용 없음>



<내용 없음>



<내용 없음>



<내용 없음>

4. Modeling

– Image classifier

- Text가 있는 Ham Image들은 글자가 산출되며, 특정 경우는 잘 인식이 되지 않는 것을 확인 할 수 있음

석나래 @ellie_daily

석촌호수에 벚꽃도 폈는데
나라도 괜찮다면 우리 같이 걸을까 아무
래도 올 봄엔 너와 견고싶다

석나래 @ellie_daily

석촌호수에 벚꽃도 폈는데
나라도 괜찮다면 우리 같이 걸을까 아무
래도 올 봄엔 너와 견고싶다

기다려도 오지않는 연락
기다리면 오는건지조차 모를때
나는 시곱바늘에 마음이 굽혀지고 있다.
흐름을만 아는것이 내 사랑을 알려가 있나
시간은 성채만 주고 떠나는것은.
그러나 눈이 없는것에게 무슨 잘못을 물으리
그것은 또 어디로가나 가서
다른 누군가의 마음을 어떻게 하겠지.

시곱바늘 / 굽은눈썹

기다려5 d 앓는 ㅈ학
iWum £능衣i奸 aiw
나는 시쌔바하 (7병이 떼f 任朴
妄◆똥만 아는간이 더 사i^w SiM-
싸名 체안 암 따따21.

그여나 놓이 枝fe곶예>(1 유 v裝 I외
고은 또 Ovmi 가서
m 궤여 마tf OmW rmi.

사쌔바 / 찻다中 I

Success is the ability to go from failure to failure
without losing your enthusiasm.

성공이란 거듭되는 실패에도 열정을 잃지 않고
계속 나아갈 수 있는 능력이다.

- 윈스턴 처칠 -

Success is the ability to go from failure to failure
without losing your enthusiasm.

성공이란 거듭되는 실패에도 열정을 잃지 않고
계속 나아갈 수 있는 능력이다.

- 윈스턴 처칠 -

너를 만나기 위해
달렸던 수 많은 시간들은
흩어지지 않고 내 주위를 맴돈다
그렇게 어둠이든 따라오면서
내가 널 놓을 생각조차 하지 못하도록
그래서 종착지인 너에게 닿기 전까지
달리는 두 다리를 절대 멈출 수 없게 한다

달리기, 서리

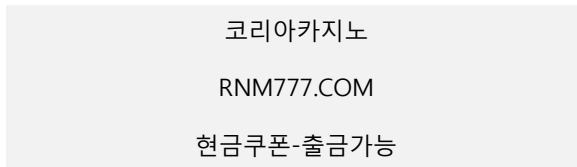
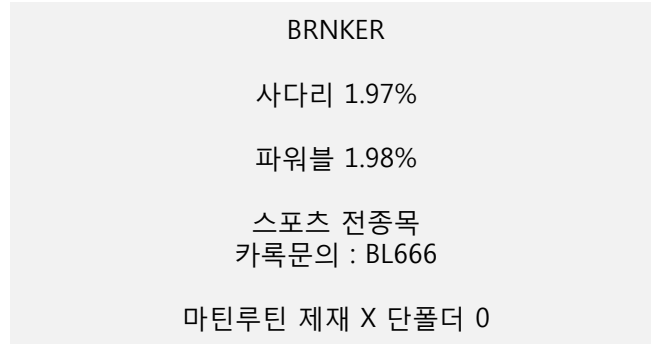
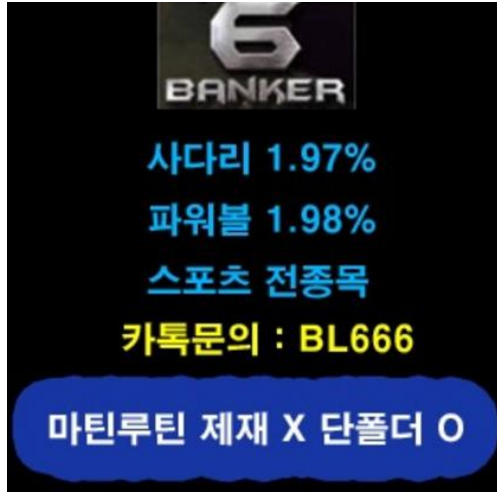
너를 만나기 위해
달렸던 수 많은 시간들은
흩어지지 않고 내 주위를 맴돈다
그렇게 어S가돈 따라오면서
내가 널 놓을 생각조차 하지 못하도록
그래서 종착지인 너에게 닿기 전까지
달리는 두 다리를 절대 멈출 수 없게 한다

달리기, 서리

4. Modeling

– Image classifier

- Text가 있는 Spam Image들은 글자가 산출되며, 특정 경우는 잘 인식이 되지 않는 것을 확인 할 수 있음



4. Modeling

– Image classifier

- Text가 있는 Spam Image들은 글자가 산출되며, 특정 경우는 잘 인식이 되지 않는 것을 확인 할 수 있음



<문자 인식이 되지 않음 T>

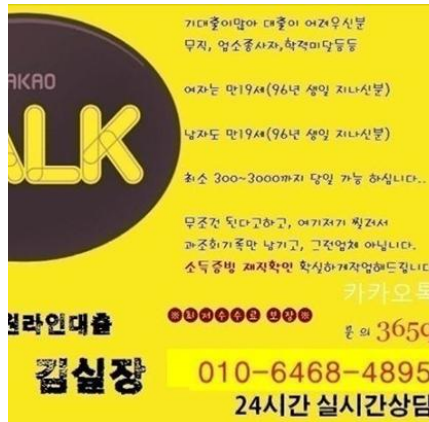


W우리카치노

HO 카지노】

첫입금시
첫충

회원가입시
원지급



7> O!OfO^ C<?O> 공시
9지, 업소f 사자,학각O•양r >

公•와는 R.^9a*(?6^ 상억 지|*•신분)

sr>E 4ri9x<(??^ 숙익 xil>xjf)

i•소 3oo~v 3ooo^7i 방이 t>-V 6에니 C, , ,

95것 수다?fr, O*7U^7i 칭건서

^7IZ?f,]꾸업회 O>님니CK
소족f 잉 JC■칭학O % , 심하w자영9칭니다.

a 망연대출

GOOOOd 00•

i^36sq

■ 名장 010-6468-4895

24시간 실시간상담



다 vat%

■방awi 항m

허혁로

며며병요수

n» 특切 atl212

4. Modeling

– Image classifier

- 배경 흰색으로 전처리
- 반전 준 경우, 그대로 흑백처리 한 경우 2가지에 대하여 동시 텍스트 수집 후 합쳐서 데이터를 생성함

〈 현재 OCR기술의 문제점 〉

1. 속도가 느림 (대량의 데이터에 적합하지 않음)
2. 필기체에 취약함
3. 배경이 존재할 경우 Edge를 잘 찾지 못하여 인식률이 떨어짐
4. 해상도가 낮을 경우 문자 추출이 잘 되지 않음

➔ 결국 인식률이 떨어지면, Text분석에 용이하지 않은 데이터가 추출이 되어 버림

	Normal	Ham	Spam
대상 이미지 개수	200	250	250
OCR 텍스트 추출 안된 경우	196	19	21

4. Modeling

– Image classifier

- 결과를 해석해 보자면, 제대로 텍스트를 불러올 수 있다면 SPAM안에 있는 term으로 충분히 분류 할 수 있다는 결론이 나옴
- 속도가 빠르고 앞서 말한 OCR의 문제점을 극복할 수 있는 알고리즘이 개발된다면 충분히 가능한 접근 방법임

	Predict(HAM)	Predict(SPAM)
HAM	249	1
NORMAL	200	0
SPAM	172	78

	Predict(HAM)	Predict(SPAM)
Actual(HAM)	449	1
Actual(SPAM)	172	78

〈 현 상황의 해결 방안 〉

1. OCR의 문제점을 극복한 새로운 알고리즘 개발
2. Text분류가 아닌 이미지 자체에 대한 classifier개발

5. Conclusion

가설 1. Instagram에서 생성한 분류기를 통해 Facebook spam 분류가 잘 작동할 것이다.

→ 아주 잘 작동하지는 않음

가설 2. Instagram 분류기와 함께 Facebook에서 만든 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

가설 3. OCR를 통해 생성한 Text를 기준으로 SPAM 이미지 분류기가 잘 작동할 것이다.

→ OCR에서 텍스트가 온전히 잘 불러와 진다면 좋은 분류기를 생성할 가능성이 있음

→ 이번 실험을 통해 기존 OCR의 문제점을 파악 하였음

→ 현 상황에서는 순수한 Image를 기반으로 한 분류기도 고려해볼 상황임 (ex) CNN output + meta input → ANN, RF, GBM..

〈 기여점 〉

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

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



THANK YOU