Journal of the Korean Institute of Industrial Engineers
Published Online, pp. 192-202, June 2017.
ISSN 1225-0988 | EISSN 2234-6457

<Original Research Paper>

인스타그램 기반의 전이학습과 게시글 메타 정보를 활용한 페이스북 스팸 게시글 판별

김준홍 · 서덕성 · 김해동 · 강필성†

고려대학교 산업경영공학부

Facebook Spam Post Filtering based on Instagram-based Transfer Learning and Meta Information of Posts

Junhong Kim · Deokseong Seo · Haedong Kim · Pilsung Kang

School of Industrial Management Engineering, Korea University

This study develops a text spam filtering system for Facebook based on two variable categories: keywords learned from Instagram and meta-information of Facebook posts. Since there is no explicit labels for spam/ham posts, we utilize hash tags in Instagram to train classification models. In addition, the filtering accuracy is enhanced by considering meta-information of Facebook posts. To verify the proposed filtering system, we conduct an empirical experiment based on a total of 1,795,067 and 761,861 Facebook and Instagram documents, respectively. Employing random forest as a base classification algorithm, experimental result shows that the proposed filtering system yield 99% and 98% in terms of filtering accuracy and F1-measure, respectively. We expect that the proposed filtering scheme can be applied other web services suffering from massive spam posts but no explicit spam labels are available.

Keywords: Spam Filtering, Facebook, Instagram, Hash Tag, Random Forest, Transfer Learning

1. 서 론

페이스북을 필두로 하여 다수의 사용자를 확보하고 있는 소셜 네트워크 서비스(Social Network Service; SNS)는 단순한 개인 간의 소통 창구 역할을 넘어 뉴스와 지식을 공유하는 새로운 미디어의 플랫폼의 역할을 수행하고 있다(Jo, 2011). 이러한 SNS에서는 정보가 불특정 다수에게 빠르게 확산된다는 특징이 있으며 최근 이를 악용하여 선정적인 문구와 사진을 곁들여 사람들의 이목을 끄는 불법 스팸(spam)이 급증하고 있는 실정이다. 이러한 불법 스팸은 SNS가 가지는 순기능의 효과를 감소시킬 뿐만 아니라 다양한 불법적인 활동(도박, 성매매 등)을 조장함으로써 사회적 문제를 야기할 수 있는 심각한 문제라고 할 수 있다.

스팸 필터링 연구는 데이터마이닝의 대표 적용 분야중 하나이며 국내에서는 2000년대부터 본격적으로 한글 스팸 필터링에 대한 연구가 수행되었다. Lee *et al.*(2011b)의 연구에서는 2,548건의 Short Message Service(SMS) 데이터를 Naive Bayesian classifier와 Support Vector Machine(SVM)을 사용하여 분류 하였고, Lee(2010)의 연구에서는 6,572개의의 이메일 학습 데이터와 3,168개의 이메일 검증 데이터를 통해 SVM과 카이제곱 통계량을 이용하여 스팸 필터링을 구현하였다. Lee *et al.*(2011a)의 연구에서는 500개의 SMS 데이터 탐색으로 생성한 특성으로 스팸 필터링을 구현 하였으며, Joe *et al.*(2009)의 연구에서는 460개 SMS 데이터를 기반으로 SVM을 이용하여 스팸 필터

이 논문은 2016년도 정부(미래창조과학부 및 교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2014R1A1A1004648, No. 2015R1A2A2A04007359, NRF-2016R1D1A1B03930729).
† 연락저자 : 강필성 교수, 02841, 서울시 성북구 안암로 145 고려대학교 산업경영공학부, Tel : 02-3290-3383, Fax : 02-929-5888, E-mail : pilsung_kang@korea.ac.kr
2016년 8월 13일 접수; 2017년 1월 2일 1차 수정본 접수; 2017년 2월 16일 2차 수정본 접수; 2017년 2월 18일 게재 확정.

링을 구현하였다.

이와 같이 국내의 스팸 필터링 연구는 이메일과 SMS 데이터 기준 10,000건 이하 비교적 소량의 데이터로 실험적으로 증명한 연구가 대부분이며 전체 SNS의 스팸탐지를 위하여 대량의 SNS의 실제 데이터를 통한 연구는 아직 없는 실정이다. 이는 일반 연구자 혹은 개인이 스팸 데이터를 다량으로 취득하기가 쉽지 않기 때문이며 현실에서 스팸 데이터가 정상 데이터에 비하여 비교적 불균형하게 분포하기 때문이다. 데이터의 취득권을 가지고 있는 페이스북 혹은 다른 여타 인터넷 운영자 입장에서도 스팸 신고 이외에는 데이터를 통하여 목표변수를 취득하기 쉽지 않다. 따라서 연구자 입장에서는 스팸 필터링 알고리즘에 대한 이해도가 높더라도 목표변수가 있는 데이터를 다량으로 수집하지 못하여 범용적인 스팸 필터링 알고리즘을 구축하기가 쉽지 않아 범용적인 스팸 필터링 알고리즘을 구축하기가 쉽지 않다.

해외에서도 스팸 필터링은 활발히 논의 되고 있는 연구 분야 중 하나이다. Zhang *et al*.(2016)의 연구에서는 1,181,735개의 트위터 계정을 포함한 Tweets 2011 dataset을 이용하였으며, 게시한 URL 정보에 근거하여 트위터 계정간의 유사도 측정에는 Shannon(2001)의 Shannon information theory을 보완한 접근법으로 그래프 기반 유저 군집을 구성하고, 트윗 횟수, 이상 URL 게재 횟수, 게재 URL 종류의 수 등을 기반으로 스패머를 정의하고 분류하는 방법을 제시하였다. Kanaris *et al*.(2006)의 연구에서는 481개의 스팸 이메일 데이터와 2,412개의 정상 이메일 데이터를 이용하여 문자를 N-gram 단위로 표현해 SVM을 사용하여 스팸 필터링을 구현하였다. Soiraya *et al*.(2012)의 연구에서는 J48 의사결정나무를 150개의 페이스북 포스트로 학습하고 75개의 테스트 포스트로 성능을 검증 하여 F1-measure 기준 약 63%의 검출력을 나타내었다. 학습에 사용된 변수로는 문자열의 길이, 평균 사용한 단어 수, 포스트의 길이, 그리고 포스트에 입력돼 있는 URL 링크의 수가 있다. 이 연구는 페이스북 포스트의 특징을 반영하는 변수를 적절히 생성했다는 장점이 있는 반면, 학습과 검증에 사용된 데이터 수가 충분하지 않고 비교적 간단한 알고리즘에 속하는 의사결정나무 한 가지만 사용했다는 한계점이 존재한다. Gao *et al*.(2012)는 187,000,000개의 페이스북 포스트와 17,000,000개의 트위터 트윗을 사용하여 개별 스팸을 탐지하는 것이 아닌 스팸 유포의 진원지를 찾는 연구를 수행하기도 하였다.

스팸성 계정을 탐색하는 연구로 Yang *et al*.(2013)은 약 1,440만개의 트위터 계정을 이용해 스팸성 계정을 탐색할 때, 그래프 기반 속성, 이웃 기반 속성, 시간 기반 속성, 자동화 기반 속성 등 4가지의 접근법을 제시하였으며, Stringhini *et al*.(2010)은 SNS 친구 신청 건수, 메시지 건수, 방문 건수 등의 로그 정보를 이용하여 스팸성 계정을 탐색하였다. 이외에도 트위터 Yang *et al*.(2011)과 시나 웨이보 Zgeng *et al*.(2015)에서 스팸을 유포하는 스패머 계정을 찾는 연구도 진행되었다. 이와 같은 스팸 유포자를 찾는 연구는 포스트의 텍스트 자체보다는 유포자 계정의 행위 패턴을 찾는데 집중한다는 특징이 있다. 따라서 스팸 유포자가 다른 사용자를 팔로우 하는 패턴, 게시글을 올리는 횟수 등을 변수로 생성한다는 점에서 텍스트를 직접 분류하고자 하는 본 연구와 다소 상이한 점이 있다고 할 수 있다.

본 연구에서는 대표적 SNS인 페이스북을 대상으로 전이학습(transfer learning)과 메타정보를 이용한 스팸 분류기를 구축하고자 한다. 본 연구에서 사용하는 페이스북 게시글은 예측 모델 구축 관점에서 다음과 같이 두 가지 문제점을 가지고 있다. 첫째, 스팸 게시글과 정상 게시글에 대한 목표 변수의 획득이 어렵기 때문에 분류 모델 구축을 위한 학습 데이터 구성이 쉽지 않다. 둘째, 스팸 게시글의 비율은 정상 게시글보다 낮기 때문에 데이터 불균형 문제가 발생하여 분류기의 정확도를 저하시키는 요인이 된다. 이러한 문제점을 해결하기 위하여 본 연구에서는 인스타그램이라는 다른 SNS에서 사용 가능한 해시태그(hash tag)를 활용하여 충분한 스팸 및 정상 게시글에 대한 목표 변수를 확보하여 스팸 분류 모델을 구축하였다. 인스타그램은 사진 위주의 SNS로서 사용자가 사진을 올리면서 직접 태그를 입력할 수 있는데 이를 해시태그라고 한다. 본 연구에서는 명백히 스팸이라 할 수 있는 해시태그로 검색하여 얻은 게시글은 검토 후 목표변수를 스팸으로 할당하였고, 스팸이 아닌 정상 게시글도 비슷한 방식으로 수집하였다. 이렇게 수집된 인스타그램의 학습 데이터를 활용하여 구축된 스팸 분류 모델을 페이스북 데이터에 적용하는 전이학습을 통해 명시적인 목표변수가 존재하지 않는 페이스북 게시글의 스팸 여부를 판별하는 분류기를 구축하였다. 이에 더하여 페이스북 게시글에서 추출할 수 있는 메타 정보들을 활용하여 분류 성능 향상을 추구하였다. 일반적으로 스팸 게시글은 문자가 표상하는 의미뿐만 아니라 게시글의 메타 정보에도 스팸 분류에 유용한 정보들을 내포하고 있기 때문이다. 예를 들어 게시글의 접촉 정보에 URL이나 '카카오톡 ID' 등이 사용되는 경우 대부분 스팸 문서로 판별될 수 있기 때문이다.

본 논문의 구성은 다음과 같다. 제 2장에서 연구의 프레임워크와 방법론에 대하여 간략히 소개하며, 제 3장에서 데이터 수집에 관한 절차와 수집 결과물을 소개한다. 제 4장에서는 스팸 필터링 시스템 구축에 필요한 데이터 전처리 절차에 대해서 설명하고 제 5장에서는 스팸 분류기 구축에 대한 부분을 서술한다. 마지막으로 제 6장에서는 결론과 함께 연구의 시사점을 논의한다.

2. 연구 프레임워크

본 연구의 전반적인 수행 절차는 <Figure 1>에 나타난 바와 같다. 먼저 인스타그램에서 스팸과 정상 해시태그를 이용하여 게시글들을 수집하고, 스팸 필터링 시스템을 적용할 페이스북 페이지로부터 역시 게시글과 함께 메타 정보를 수집한다. 수

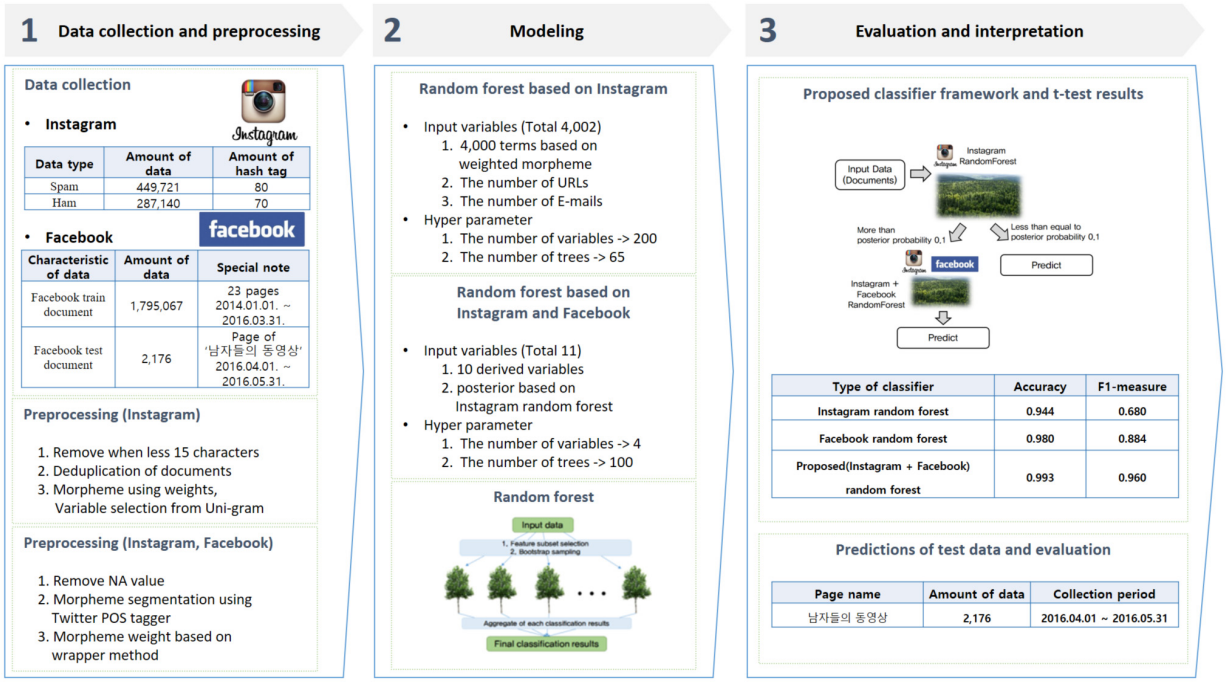

Figure 1. Research Framework

집된 데이터에 대하여 인스타그램/페이스북에 따라 적절한 전처리를 수행하였다. 두 번째 단계에서는 인스타그램의 텍스트 데이터를 이용하여 스팸을 분류하는 모델과 인스타그램과 페이스북의 메타 정보를 함께 활용하여 스팸을 분류하는 모델을 랜덤포레스트를 이용하여 학습하였다. 두 가지의 랜덤포레스트를 이용한 이유는 특정 게시글의 경우 메타정보를 이용하지 않고 텍스트만 활용하더라도 탐지가 잘 되며, 이러한 게시글들을 전자의 모델을 통해 1차적으로 필터링을 한 뒤, 나머지 게시글들에 대해서 메타정보까지 활용한 모델을 통해 보다 정확한 필터링을 하기 위해서이다. 마지막 단계에서는 학습한 모델을 기반으로 50회 반복실험을 수행한 뒤 여섯 가지 평가지표를 통해 제안된 방법론의 성능을 평가하였다.

본 연구에서 적용한 분류기는 의사결정나무 기반의 앙상블 기법인 랜덤포레스트이다. 랜덤포레스트는 배깅(bagging) 기법을 이용하여 서로 다른 학습 데이터 부트스트랩을 생성하며, 개별 의사결정나무의 분기점 선택이 임의의 적은 변수들만을 고려한다. 이 과정을 통해 랜덤포레스트에 속한 의사결정나무들은 두 가지의 다양성을 확보할 수 있으며 이를 조합하여 좋은 성능을 갖는 하나의 분류기를 생성할 수 있다(Breiman, 2001). 랜덤포레스트는 특히 노이즈에 대해 강건한 특징을 가지며 분류 문제에 있어 기존에 존재하는 179가지 분류 알고리즘을 121개의 데이터셋에 대하여 성능을 비교해본 실증적 연구에서도 가장 우수한 분류 성능을 나타내는 것으로 확인되었다(Fernández-Delgado *et al.*, 2014). 이에 본 연구에서는 분류 성능이 좋은 랜덤포레스트를 스팸을 분류하는 분류기로 적용하였다.

또한 본 연구에서는 인스타그램의 스팸/정상 표식을 이용하여 게시글을 학습한 분류기로 페이스북의 게시글의 스팸 여부를 예측하였다. 이는 SNS의 종류와는 무관하게 스팸성 게시글의 텍스트 사용 패턴이 비슷한 것이라는 가정을 바탕으로 수행한 것이다. 일반적으로 기계학습과 데이터마이닝에서는 모델 구축과 활용에 사용하는 데이터가 동질하다는 가정을 한다. 그러나 현실에서는 이러한 가정을 지키기 어려운 상황이 발생한다. 이러한 상황에서 양질의 데이터를 충분히 구하기 힘들 경우, 현실적인 대안으로 비슷한 영역의 다른 데이터를 차용해 사용하는 전이학습(transfer learning) 방법을 사용한다(Pan and Yang, 2010). 본 연구에서는 이러한 관점에서 상대적으로 표식 데이터의 수집이 수월한 인스타그램의 게시글을 학습하여 페이스북의 스팸 게시글을 분류하는 모델을 구축하고자 한다.

3. 데이터 수집

페이스북의 스팸 게시글을 필터링하기 위해 1차적으로 스팸/정상 게시글을 명확히 구분할 수 있는 인스타그램 텍스트를 수집하였다. 총 80개의 정상 해시태그와 70개의 스팸 해시태그에 대해 게시글들을 수집한 결과 정상 게시글은 449,721건, 스팸 게시글은 287,140건이 수집되었다. 또한 해당 스팸 필터링 시스템을 페이스북에 적용하기 위하여 총 23개의 페이스북 페이지에서 2014년 1월 1일부터 2016년 3월 31일까지 1,795,067건의 게시글 및 관련된 파생 변수를 수집하였다. 최종적으로 구축된 스팸 필터링 시스템의 성능을 확인하기 위한 검증 데이터로는 "남자들의 동영상" 페이지의 학습 데이터 기준 이후인 2016년 4월 1일부터 2016년 5월 31일까지 페이지에 게시된 2,176건

의 게시글을 사용하였다. “남자들의 동영상” 페이지를 검증용 데이터로 선정한 이유는 다른 페이지들에 비해 비교적 스팸성 글의 비중이 비교적 높으며 다양한 스팸이 게시되고 있다고 판단되었기 때문이다. <Table 1>은 수집한 데이터 이름과 개수 그리고 특이사항이다.

Table 1. Raw Data Collection

Name of data	No. of posts	Notes
Instagram HAM data	449,721	80 Hash tag
Instagram SPAM data	287,140	70 Hash tag
Facebook data (Training data)	1,795,067	23 page 2014. 01. 01.~2016. 03. 31.
Facebook data (Test data)	2,176	Page of ‘남자들의 동영상’ 2016. 04. 01.~2016. 05. 31.

3.1 인스타그램 데이터 수집

인스타그램 데이터는 해시태그 기준으로 수집하였다. 정상 혹은 스팸성 게시글이 대다수인 해시태그에 대하여 검색을 통해 데이터를 수집하는 절차를 진행하였다. 이를 위해 인스타그램에서 수집한 80개의 정상 해시태그와 70개의 스팸 해시태그 및 각 해시태그별로 데이터 수집을 하였다. 정상의 경우 여러 분야를 모두 고려하도록 수집되었으며, 스팸의 경우 도박, 성인물, 대출 관련 정보 등 여러 주제를 고려하여 수집하였다. 스팸성 해시태그의 경우는 논문에 서술하기에는 부적절한 단어가 많이 포함되어 있어 목록은 제외하였다. 만약, 해당 정보가 필요할 경우 1저자에게 이메일로 요청하면 제공하도록 하겠다.

인스타그램 데이터의 특징은 다음과 같이 두 가지가 있다. 첫째, 해시태그가 있다는 점이다. 이를 통해 다양한 해시태그들을 기준으로 여러 분야의 게시글을 수집하는 것이 가능하게 되었다. 페이스북 혹은 비슷한 SNS에서 정상과 스팸은 매우 불균형하게 존재하므로 대량의 스팸 게시글을 손쉽게 수집하지 못하는 한계가 있다. 하지만 인스타그램에서는 해시태그 단위로 게시글 수집이 가능하므로 비교적 수월하게 균형적인 스팸 게시글과 정상 게시글을 수집할 수 있다는 장점이 있다. 또한 스팸 게시글을 수집할 시 도박, 성인물, 대출 등 원하는 분야의 데이터를 수집 가능하다는 점은 텍스트를 이용한 스팸 분류기 구축 과정에서 장점으로 작용한다. 둘째, 데이터의 대하여 목표변수를 비교적 손쉽게 설정 가능하다는 것이다. 예를 들어 “#명언”이라고 검색했을 경우 5,996개의 수집된 게시글 중 대부분이 정상이라는 것을 쉽게 확인할 수 있으며, “#사다리픽”이라고 검색했을 경우 5,591개의 수집된 게시글의 대부분은 스팸 게시글 이라는 것을 쉽게 확인할 수 있다. 이를 기반으로 인스타그램 게시글에 대한 표식을 설정하여 교사 학습을 진행 할 수 있다.

3.2 페이스북 데이터 수집

본 연구에서는 구독자 순과 스팸성 게시글/댓글이 존재할 가능성을 고려하여 대상 페이스북 페이지들을 선정하였다. <Table 2>에 서술한 1~10번의 페이지는 ‘socialbakers’(https://www.socialbakers.com/statistics/facebook/pages/total/south-korea/)에서 제공하는 대한민국 Facebook 페이지 순위 중 연예인 페이지를 제외한 좋아요 기준 상위에 있는 10개 페이지이며 11~23번의 페이지는 스팸성 게시글/댓글 수집을 위한 13개 등 총 23개의 페이지를 최종적으로 선정하여 데이터를 수집하였다.

4. 데이터 전처리

4.1 1차 전처리 & 목표변수 수정

인스타그램에서 수집한 데이터를 바탕으로 다음과 같이 1차 전처리를 행하였다. 첫째, 문자열 길이가 15 이하인 게시글을 제거하였다. 이는 인스타그램 특성상 상당히 짧은 글들이 존재 할 수 있는데 이러한 게시글들은 문서적 특성을 가지기 힘든 경우가 많기 때문이다. 둘째, 데이터를 수집할 때 예상치 못한 이유로 특정 항목에 대해 정보가 수집되지 않는 경우가 발생하는데 본 연구에서는 모든 메타 정보가 존재하는 게시글

Table 2. Facebook Pages Considered in this Study

Page Name		
1. 세상에서 가장 웃긴 동영상	9. 메갈리아4	17. 남자들의 축구
2. 남자들의 동영상	10. 레진코믹스	18. 세상에서 가장 소름돋는 라이브
3. 여행에 미치다	11. 잡지 사심	19. 19세 이상만
4. 네임드사다리	12. 한국성폭력상담소	20. 중고차 전국출장매입
5. 남자의 관리	13. 민주노총	21. 도우미론
6. 남자옷덕후	14. 스포츠 마니아 일루모여	22. 사다리전문페이지
7. 바카라 놀이터 네임드사다리	15. 여자들의 동영상	23. 대학내일
8. 사다리프젝전문	16. 안전놀이터/네임드사다리 bs-sky-1.com 안전토토	

Table 3. Number of Posts before and after Preprocessing

	Raw Data		After duplicated post removal		After target modification	
	HAM	SPAM	HAM	SPAM	HAM	SPAM
No. of posts	449,721	287,140	178,765	20,530	181,590	17,705

들을 분석 대상으로 삼기 위하여 이러한 게시글들은 제거하였다. 셋째, 중복된 게시글은 한번만 사용하도록 전처리를 수행하였다. 인스타그램은 한 게시글에 대해 여러 개의 해시태그를 허용하기 때문에 여러 해시태그의 검색 결과 하나의 게시글이 반복해서 결과물로 반환될 수 있다. 또한 일반적인 SNS의 특성으로 인해 동일 사용자가 한 게시글을 여러 번 반복하여 게시하거나, 다른 사용자들에 의해 원 게시글이 복제되어 중복되는 경우 또한 존재한다. 따라서 이렇게 중복된 게시글들이 발생할 경우 하나의 게시글만 남겨두고 모두 삭제하였다. 넷째, 해시태그 기반으로 표식을 생성할 경우 발생할 수 있는 잠재적 오류를 방지하기 위하여 20,530건의 스팸 게시글에 대해서는 전수 검수를 통해 표식을 수정하였다.

<Table 3>은 인스타그램 데이터에 대한 전처리를 모두 수행한 후 정상 게시글과 스팸 게시글의 수를 나타낸 것이다. 스팸 게시글의 경우 중복이 정상 데이터에 비하여서 상대적으로 빈번하게 발생하여 수집된 데이터 기준 10% 미만의 게시글만이 최종적으로 남아있는 것을 알 수 있다. 데이터의 개수가 약 20,530개의 데이터로 감소 한 것을 알 수 있다. 또한 표식을 수정함으로써 약 15%의 스팸 게시글이 추가적으로 정상 게시글로 이동하여 최종적으로 181,590건의 정상 게시글과 17,750건의 스팸 게시글이 확보되었다.

4.2 형태소분석(POS tag) & 인스타그램 기반 단어 변수선택

본 연구에 사용된 게시글들은 SNS에서 사용되는 용어 및 표현 방식이 만연한 관계로 표준어의 문법을 따르지 않는 경우가 많다. 따라서 이러한 특징을 보다 정확하게 반영하는 스팸 분류기를 구축하기 위하여 표준어 기반의 형태소 분석기가 아닌 트위터 게시글 기반의 형태소 분석기를 사용하여 형태소 분석을 수행하였다.

Table 4. Contingency Table for Assignining Supervised Term weight

Feature \ Category	C	$\overline{C}$
F_i	$F_i\,C$	$F_i\,\overline{C}$
$\overline{F_i}$	$\overline{F_i}\,C$	$\overline{F_i}\,\overline{C}$

다음으로, 게시글의 스팸 여부를 분류하기 위해 유니그램(uni-gram)기반의 교사적 용어 가중치(supervised term weight) 기반의 변수 선택기법을 사용하였다. 사용한 방법은 개별 용어들이 등장한 문서 빈도를 이용하여 계산하는 엔트로피 기반의 Information gain과 Prob weight 두 가지이다(Quan *et al*., 2011). 두 지표를 산출하기 위한 기초 자료는 <Table 4>와 같으며 엔트로피와 Information gain, Prob weight를 산출하는 식은 각각 Eq. (1), Eq. (2)과 Eq. (3)에 나타나 있다.

$$\begin{aligned} &Entropy(given\,F_i) \\ &= P(F_i)[\textstyle\sum_{j\in P,N} - P(C_j|F_i)\times\log(P(C_j|F_i))] \\ &+ P(\overline{F_i})[\textstyle\sum_{j\in P,N} - P(C_j|\overline{F_i})\times\log(P(C_j|\overline{F_i}))] \end{aligned} \tag{1}$$

$$\begin{aligned} &Information\,gain \\ &= Entropy(absent) - Entropy(given\,F_i) \end{aligned} \tag{2}$$

$$\mathrm{Prob}\,weight = \log\left(1+\frac{F_iC}{F_i\overline{C}}\,\frac{F_iC}{\overline{F_i}C}\right) \tag{3}$$

F_i와 $\overline{F_i}$는 특정 'i'번째 uni-gram 단어의 출현 유무를 의미하며, C와 $\overline{C}$는 특정 범주(class) 혹은 해당 범주가 아닌 경우를 의미한다. 본 연구에서는 C가 스팸을 의미하며 $\overline{C}$가 정상을 의미한다. $F_i\,C$는 스팸 범주에서 F_i 단어가 있는 문서 수, $F_i\,\overline{C}$는 정상 범주에서 F_i 단어가 있는 문서 수, $\overline{F_i}\,C$는 스팸 범주에서 F_i 단어가 없는 문서 수, $\overline{F_i}\,\overline{C}$는 정상 범주에서 F_i 단어가 없는 문서 수를 의미한다. 따라서 $F_i\,C+F_i\,\overline{C}+\overline{F_i}\,C+\overline{F_i}\,\overline{C}$는 전체 문서의 수와 같다. Eq. (1)에서의 'j'의 'P'와 'N'은 Positive와 Negative를 즉, 스팸과 정상을 의미한다. Prob weight와 엔트로피 기반의 Information gain 두 방법 모두 특정 유니그램 단어 변수가 본 연구에서의 스팸, 정상 범주의 분류를 위해 효과적이라는 기저는 일맥상통하나, Prob weight의 경우 스팸 범주 기반의 방식이며 엔트로피는 스팸과 정상범주를 동시에 반영한 방식인 것을 수식을 통해 알 수 있다.

본 연구에서 수집한 인스타그램 게시글의 전체 집합(corpus)에서 트위터 형태소 분석기로 산출된 형태소는 총 144,187개이다. 해당 형태소를 모두 사용하는 것은 계산복잡도 면에서 비효율적이며 성능도 좋지 않을 확률이 높다. 본 연구에서는 사용할 형태소는 트위터 형태소 분석기 기준 형용사(Adjective), 부사(Adverb), 알파벳(Alpha), 외래어(Foreign), 한글 입자(Korean particle), 명사(Noun), 동사(Verb), 알 수 없음(NA)의 총 여덟 가지를 사용하였다. 이 중 가중치 기준 상위 4,000개를 최종적으로 선택하여 개별 게시글을 벡터 형태로 표현하는데 사용하였다. <Figure 2>는 각 형태소별로 Prob Weight와 Entropy 기반의 Information Gain 값을 정렬한 결과로써 초기 수천 개의 단어 이후에는 각 지표의 값이 거의 0의 값에 수렴(정상과 스팸 분류에 영향을 미치지 않음)하는 것을 알 수 있다. 따라 본 연구에서는 상위 4,000개의 형태소를 바탕으로 Bag-of-word 방식을 사용하여 각 게시글을 4,000차원의 벡터로 변환하였다.

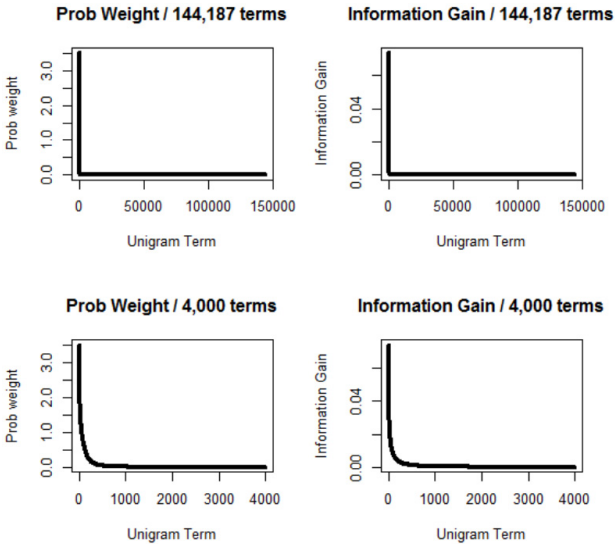

Figure 2. Each Uni-gram Word Weight Graph by Probability weight and Information Gain(Top : Total 144,187 terms)/(Bottom : Total 4,000 terms based 8 Selected Morpheme)

4.3 Facebook 기반의 파생변수 설정

본 연구에서는 인스타그램의 단어 기반 변수 이외에도 데이터 탐색을 통해 페이스북 스팸 분류에 적합한 파생변수를 추가 생성하였다. 그 이유는 스팸 게시글의 경우 정상 게시글과는 확연히 다른 패턴을 보이는 경우가 존재하며 이는 게시글의 메타 정보를 통해 확인할 수 있는 경우도 존재하기 때문이다. 본 연구에서 사용한 페이스북 게시글의 파생 변수명과 그에 대한 설명은 <Table 5>에 나타나 있다.

'Enter_Per_String' 변수는 일반적으로 스팸 데이터의 경우 일반 댓글과 포스팅의 글보다 엔터의 수가 많다는 특징을 가지고 있다. 이를 스트링의 전체 길이로 나누어 줌으로써 가변적인 길이에 대한 상황을 보정하였다. 'Num_Alpha_Number' 변수는 연속적으로 '알파벳+숫자'인 단어의 개수를 측정한 것이다. 최근 페이스북의 스팸에서는 URL로 홍보할 뿐만 아니라, '카카오톡' 등의 ID로 홍보를 하고 있는 것을 데이터 탐색을 통해 알 수 있다. 따라서 ID는 숫자부터 시작하는 경우가 거의 없기 때문에 문장의 형태소 중에서 '알파벳+숫자'가 연속적으로 나오는 경우를 측정 하였다. 이는 ID로 간주하기 때문이다. 'Num_Number_Term_Per' 변수는 문서 안에 숫자의 개수를 측정하고 스트링 길이로 나눠준 것이다. 스팸성 문장은 일반 문장보다 숫자가 나올 확률이 높기 때문이다. 'Target_BY_Spam', 'Target_BY_Spam_Per' 변수는 인스타그램의 선택된 단어로 학습된 랜덤포레스트를 통해 산출된 사후 확률을 다시 입력 변수로 설정하였다. 스팸성 단어가 있는 경우, 인스타그램 기반의 스팸 필터링 법칙을 적용 할 수 있기 때문이다. 'Num_Alpha_Term_Per' 변수는 문장의 알파벳의 개수를 측정하고 스트링 길이로 나눠준 것이다. 스팸성 게시글이 일반 게시글보다 알파벳 단어가 많이 작성될 확률이 높은 것을 바탕으로 파생 되었다. 'Num_of_Change_Alphabet' 변수는 페이스북의 댓글에 자신의 친구 혹은 지인을 태그 입력 할 수 있는 기능이 존재하는데 상당수의 정상 댓글에서 이러한 태그를 사용하는 점에서 고안된 변수이다. 보통 한국인의 이름이 영어로 저장되어 있을 경우 홍길동의 경우 'Gil Dong Hong'처럼 첫 글자가 대문자이고 다음 글자부터가 소문자로 저장이 되는 구조이며 고안한 파생 변수를 통해 이러한 한글 사용자명을 탐지할 수 있게 된다. 'Num_Puntu_Term_Per', 'Num_Punc_star 변수는 한자기반의 특수문자 개수(예 ※, ★, ◆ 등)와 쉬프트기반 특수문자(예 !, @, #, $ 등) 개수를 스트링 길이로 나누어준 것이며, 'Start_Punct_Hanja', 'Start_Punct_punct' 변수는 한자기반과 쉬프트 기반의 특수문자로 문서가 시작하는지에 관한 시작유무에 관한 변수이다.

5. 스팸 분류 모델 구축

본 연구에서는 스팸 게시글을 탐지하기 위한 문서 분류 알고리즘으로 랜덤포레스트를 사용하였다. 랜덤포레스트는 범용적으로 성능이 좋은 모델로 알려져 있는데 이는 Fernández-Delgado *et al.*(2014)의 연구에서 다시 한 번 실험적으로 검증

Table 5. Derived Variables for Facebook Text Spam Filtering

NO	Variable name	Description
1	Enter_Per_String	Number of enter keys/String length
2	Num_Alpha_Number	Number of alphabets and numeric values
3	Num_Number_Term_Per	Number of numeric values/String length
4	Target_BY_Spam	Spam probability predicted based on Instagram data
5	Num_Alpha_Term_Per	Number of alphabets/String length
6	Target_BY_Spam_Per	Spam probability predicted based on Instagram data/Strings length
7	Num_of_Change_Alphabet	Number of capital letters
8	Num_Puntu_Term_Per	Number of Chinese special characters/String length
9	Num_Punc_star_Term_Per	Number of special characters typed with the shift key/String length
10	Start_Punct_Hanja	Indication whether the post begins with a chinese special character
11	Start_Punct_Punct	Indication whether the post begins with a special character typed with the shift key

되었다. 이 연구에서 총 121개의 UCI 분류 데이터 셋에 대해 179가지의 분류 알고리즘을 이용하여 분류 성능을 비교한 결과 랜덤포레스트관련 알고리즘이 비교적 다른 알고리즘에 비해 우수한 분류 성능을 나타내는 것으로 확인되었으며, Oh *et al.* (2014)의 연구에서도 텍스트뿐만 아니라 야구 데이터에서도 랜덤 포레스트가 일반적으로 우수한 성능을 보여줌을 실증적으로 확인 하였다. 또한 랜덤포레스트는 앙상블 기반의 모델 중 순서가 필요 없는 Implicit Ensemble 기반의 알고리즘으로 분류기들을 동시에 병렬적으로 생성하는 것이 가능하므로 Explicit Ensemble 기반의 부스팅(boosting) 계열 기법에 비해 시간적으로 효율적인 모델 구축이 가능하다는 장점이 있어 본 연구에서는 랜덤포레스트를 분류기로 이용하였다.

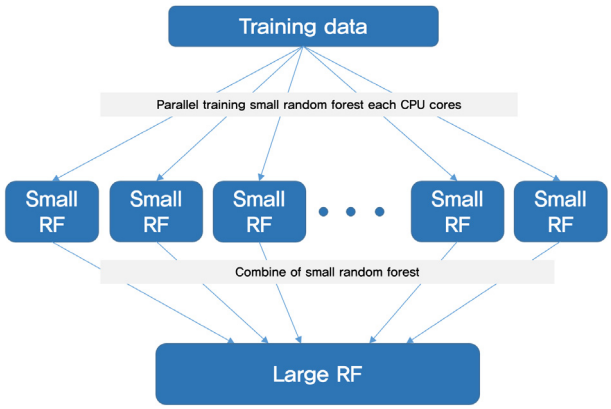

Figure 3. Random Forest Parallel Learning Framework

랜덤포레스트 기반의 스팸 분류기는 <Table 6>에 나타난 바와 같이 Prob weight 및 Information Gain 기반의 상위 4,000개의 형태소 변수의 각 문서별 단어 빈도, URL의 개수 및 Email 문자열의 개수 등 총 4,002가지의 변수를 사용하여 구축되었다. 또한 랜덤포레스트는 앙상블의 다양성을 확보하기 위해 원래 입력 변수 중 일부분을 선택하여 의사결정나무 분기에 사용하는데, 본 연구에서는 분기마다 200개 변수를 무작위로 선택하도록 설정하였으며 앙상블을 구성하는 단일 의사결정나무의 수는 65개로 설정하였다.

본 연구에서 사용된 학습 데이터는 정상 게시글이 181,590건, 스팸 게시글이 17,705건으로 스팸 게시글의 비율이 상대적으로 낮은 불균형한 형태를 지니고 있다. 분류 알고리즘인 랜덤포레스트는 범주간 불균형이 존재하는 경우 모델이 민감하게 반응하여 일반화가 잘 되지 않는다는 특징을 가지고 있다. 본 연구에서는 불균형 문제로 인한 분류기의 일반화 성능을 산출하기 위해 다수의 범주인 정상 게시글 중 25,000건을 무작위 추출한 데이터와 스팸 게시글 17,705건을 사용하여 각 의사결정나무마다 학습에 사용하였다.

두 가지 방식의 형태소 선택 기법을 이용한 랜덤포레스트에 대해 붓스트랩 생성에 사용되지 않은 데이터(out-of-bag data)를 사용한 검증 정확도는 <Table 7>에 나타난 바와 같다. Sensitivity는 스팸 게시글 중 모델에 의해 탐지된 스팸 게시글의 비율이며 Precision은 모델이 스팸이라고 판별한 게시글 중 진짜 스팸 게시글의 비율이다. Specificity는 정상 게시글 중 모델에 의해 판별된 정상 게시글의 비율, Accuracy는 전체 게시글 중 스팸/정상으로 정분류된 게시글의 비율이다. BCR은 스팸과 정상 게시글 정확도(Sensitivity와 Specificity)의 기하평균이고 F1은 Sensitivity와 Precision의 조화평균이다. 검증 정확도의 관점에서 보면 정상 게시글을 판별하는 정확도는 크게 차이가 나지 않으나 Prob Weight가 Information Gain에 비하여 스팸 게시글을 60건 정도 더 정확하게 탐지함으로써 Precision이 다소 높게 나타나는 것을 알 수 있다. 따라서 본 연구에서는 최종적으로 Prob Weight 방식을 이용한 형태소 선택을 통해 변수를 구축하여 페이스북 게시글 스팸 탐지에 적용하기로 하였다. 그 이유는 첫 번째로, 오분류 비용을 고려할 때 스팸을 정상으로 예측하는 것이 그 반대의 상황보다 손실비용이 더 높기 때문이며, 두 번째로, Prob Weight는 스팸의 단어의 비중으로 만들어진 분류기이지만 Information Gain은 정상과 스팸에 두 가지 범주의 비중으로 생성된 단어 집합이기 때문이다. 실제로 페이스북에서 사용하는 정상단어와 인스타그램에서 사용하는 정상단어가 다르게 쓰일 가능성이 있으며, 스팸단어의 비중을 둔 Prob Weight 분류기가 페이스북에 더 잘 작동할 가능성이 높기 때문이다.

Table 6. Input Variables for Construct Random Forset(Number of variables)

	Input variable based on Information gain	Input variable based on Prob weight
Uni-gram morpheme variable	Top 4,000 morpheme variable based on Information gain(4,000)	Top 4,000 morpheme variable based on probability weight(4,000)
Additional variable 1	The number of URL(1)	The number of URL(1)
Additional variable 2	The number of E-mail(1)	The number of E-mail(1)
Number of total variables	4,002	4,002

Table 7. Out-of-Bag(OOB) Error of Two Spam Filtering Models

Term selection method	Sensitivity	Precision	Specificity	Accuracy	BCR	F1
Information Gain	0.986	0.968	0.977	0.981	0.981	0.977
Prob Weight	0.987	0.971	0.979	0.982	0.983	0.979

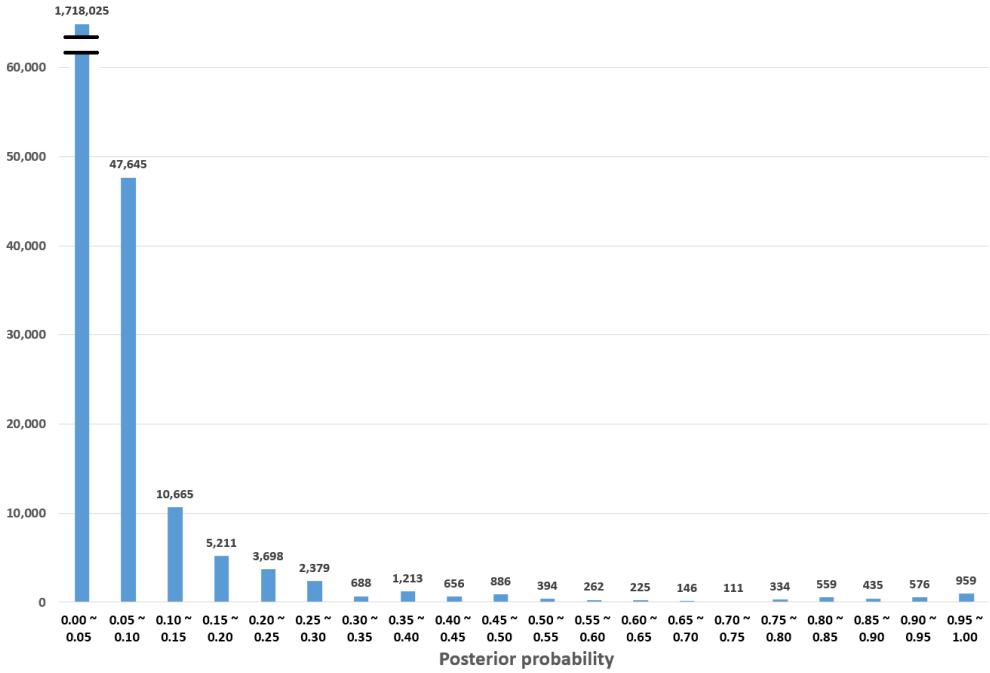

Figure 4. Random Forest Posterior based on Facebook Documents

Table 8. Example of Facebook Text Spam Classification Results from Instagram-based Spam Classifier(Spam Probability)

Example of false positive	Example of false negative	Example of true positive
도박 마약ㅌㅋㅋㅋㅋㅋㅋ짱웃갸 **(1.000)**	2개 무료 증정 분석깅 PD80**(0.123)**	매일 단톡방에서 진행되는 사다리프젝. 인원이 다 차면 모집을 안합니다. 선착순 30명만 받습니다. 100% 무료픽 단톡방입니다. 카톡 star345**(0.831)**
대출해서 밥 사줘야 할 거 같다. **(0.923)**	안녕하세요^^ :) 페이스북 자동 댓글 프로그램 임대!! #시간당 100개 이상 #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% 믿을 수 있는 출장대행 홈피주소 www.kiSS343.Com 섹시한 여대생 대기중입니다**(0.723)**

<Figure 4>는 인스타그램 데이터를 바탕으로 학습된 랜덤포레스트 기반의 스팸 분류기를 이용하여 추정된 1,795,067개의 페이스북 게시글의 스팸 확률의 일원빈도 그래프이다. 대부분의 페이스북의 게시글들은 정상 게시글에 가까운 것을 알 수 있으며, 스팸 확률 0.1 이하의 문서를 무작위 추출하여 확인한 결과, 스팸성 게시글은 한 건도 존재하지 않았다. 스팸 확률이 0.1 초과인 페이스북 게시글 수는 29,397개로 스팸 확률 0.1 이하인 문서에 비하여 비교적 소량인 것을 확인 할 수 있다. 사후 확률 0.1 초과의 데이터에서는 정상과 스팸이 섞여 있는 것을 확인 할 수 있다. 모든 게시글에 대하여 정상/스팸 여부를 사람이 판별한 결과, 스팸성 게시글은 총 2,729건, 정상 게시글은 총 26,668건으로 판별되었다. 이를 인스타그램 Prob weight 기준 선택된 단어 학습 기반의 랜덤포레스트만을 사용하여 예측할 경우의 정분류 및 오분류 예시는 <Table 8>에 나타난 것과 같다. 'Example of false positive'는 실제 정상의 문장이지만 스팸으로 사후 확률이 높게 예측된 경우를 의미한다. 의미를 해석해 보면, '대출', '토토', '도박', '마약' 같은 단어들이 짧은 글 안에 포함됨으로서 오분류가 될 확률이 높다고 생각할 수 있다. 'Example of false negative'는 페이스북에서 새롭게 나타난 스팸 주제인 '댓글 프로그램'과 '분석기'를 '분석깅'으로 회피하였다. 또한, 작성 문장이 상당히 짧아, 'Prob weight 기반 단어 백터 공간상에서 모든 변수의 값이 0으로 생성될 확률이 높으므로, 사후 확률이 낮게 측정되는 것을 확인 할 수 있다. 따라서 본 연구에서는 사전에 <Table 5>에서 제안한 파생변수를 함께 사용하는 페이스북 스팸 필터링 분류기를 구축하였다.

Table 9. Spam Filtering Performance of Three Random Forests(Values without/with parenthesis are the average/standard deviation of 50 trials, respectively)

	Sensitivity	Precision	Specificity	Accuracy	BCR	F1
Instagram RF	0.728 (0.037)	0.641 (0.038)	0.963 (0.004)	0.944 (0.005)	0.837 (0.021)	0.680 (0.029)
Facebook RF	0.953 (0.017)	0.824 (0.027)	0.982 (0.003)	0.980 (0.003)	0.967 (0.008)	0.884 (0.018)
Instagram+Facebook RF	0.963 (0.015)	0.958 (0.014)	0.996 (0.001)	0.993 (0.002)	0.979 (0.008)	0.960 (0.009)

Table 10. P-values for Two Hypotheses on Spam Filtering Performance

Hypothesis \ Valid index	Sensitivity	Precision	Specificity	Accuracy	BCR	F1
Instagram RF = Instagram+Facebook RF	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Facebook RF = Instagram+Facebook RF	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

<Table 9>는 인스타그램 문서에서 Prob weight를 기준으로 선택된 단어 기반의 랜덤포레스트(이하; 인스타그램 RF)와 <Table 6>에서 서술한 11가지의 파생 변수 중 인스타그램 단어 정보인 'Target_BY_Spam' 변수와 'Target_BY_Spam_Per' 변수를 제거한 순수 페이스북에서 생성 할 수 있는 파생변수만을 고려한 랜덤포레스트(이하; 페이스북 파생변수 RF), 인스타그램의 단어 기반 정보와 페이스북의 특성에 맞춘 파생변수 11가지를 모두 변수로 구축한 랜덤포레스트(이하; 인스타그램+페이스북 파생변수 RF) 총 3가지 모델을 사용하여 인스타그램의 단어기반으로 생성된 랜덤포레스트의 예측 사후 확률 0.1 초과 경우인 29,397개 중 27,397개를 학습용 데이터로 2,000개를 검증용 데이터로 50번 반복 실험한 결과이다. 페이스북 파생변수를 사용한 2가지의 랜덤포레스트(페이스북 파생변수 RF, 인스타그램+페이스북 파생변수 RF)는 나무의 개수로 100개, 하나의 나무에서 사용한 변수는 4개를 이용하였다.

재미있는 결과는 페이스북에서 생성한 파생변수로만 사용한 랜덤포레스트가 인스타그램 문서기반의 랜덤포레스트 분류기에 비하여 모든 지표에서 성능이 비교적 우수하게 나타났다는 것이다. 이는 페이스북의 스팸게시물 구조가 정상게시물 구조와 비교하여 보았을 때 생성된 파생변수의 특성과 상이하게 다르게 분포하고 있다는 것을 의미한다. 인스타그램의 문서기반 랜덤포레스트를 사용한 결과는 정확도 측면에서는 나쁘지 않은 성능을 보이지만 목표변수 중 중요한 스팸분류기의 평가지표인 F1-Measure에서는 상대적으로 낮은 약 68%의 성능을 보였다. 다음으로는 페이스북의 파생 변수만을 사용한 결과에서는 단순 정분류율 기준 약 98%, F1-measure 기준 약 88%의 정확도를 나타내었다. 본 연구에서 제안한 인스타그램에서 추출된 학습된 단어와 페이스북의 파생변수를 모두 사용한 랜덤포레스트는 단순 정확도 기준 약 99% F1-measure 기준 약 96%으로 가장 우수한 탐지 정확도를 나타내는 것을 확인하였다. 또한 여섯 가지 지표에 대하여 'Instagram RF'와 Instagram+Facebook RF' 그리고 'Facebook RF'와 'Instagram+Facebook RF'에 대하여 예측 성능 차이에 대한 통계적인 유의성 검증을 시행한 결과 <Table 10>과 같이 모두 p-value가 0.01 미만으로 산출되어 제안하는 방법론이 다른 방법론에 비해 유의미한 탐지 정확도의 향상을 이루어 냈음을 알 수 있다.

현재까지 분류기 검증에 사용된 데이터는 2014년 01월 01일부터 2016년 03월 31일까지의 페이스북 데이터이므로 생성된 최종 분류기의 테스트 데이터의 성능을 평가하기 위하여 2016년 4월 1일 부터 2016년 5월 31일까지 페이스북 '남자들의 동영상' 페이지에 게시된 2,176건의 게시글 및 댓글에 대하여 스팸 분류기를 적용하였으며, 그 과정은 <Figure 5>에 나타난 것과 같다.

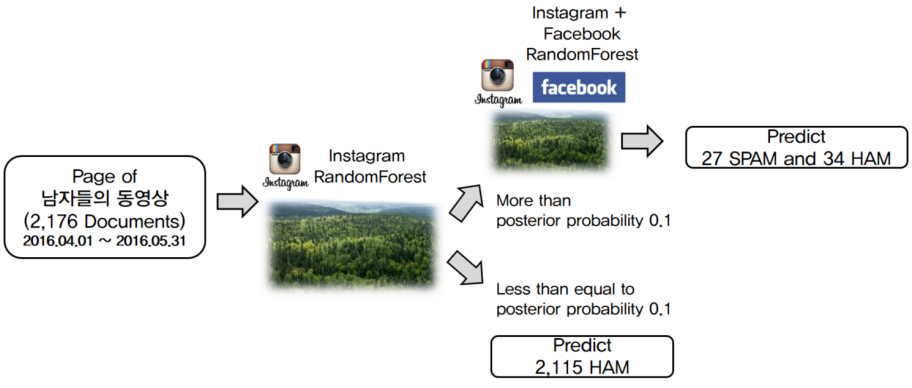

Figure 5. Test Data Classification Framework and Prediction Results

먼저, 인스타그램 기반의 랜덤포레스트로 사후 확률이 0.1 이하인 경우가 2,176건 중 2,115건이며 해당 문서는 모두 정상인 것을 검증 하였다. 사후확률 0.1초과의 경우는 61개로 확인되었으며 61개에 대하여 페이스북과 인스타그램의 데이터를 동시에 사용한 랜덤포레스트를 사용하여 분류한 결과 27개를 스팸으로 분류 하였다. 해당 분류된 27개의 문서는 모두 스팸인 것을 확인하였다. 정상 게시물로 예측한 37개중에서는 유일하게 1개가 스팸이었으며 결론적으로 전체 테스트 데이터 2,176건 중 1건을 오분류하는 결과를 산출하였다.

테스트 데이터에 대한 혼동행렬과 각 평가 지표에 대한 값들은 <Table 11>과 <Table12>에 나타난 것과 같다. 본 연구에서 제안하는 인스타그램의 텍스트 정보와 페이스북의 메타 정보를 사용하여 만든 스팸 필터링 시스템은 테스트 데이터 기준 F1-measure기준 98.2%, 정분류율 기준 99.9%의 탐지 정확도를 나타내었다. 이는 실제 페이지에 사용할 만큼의 성능이라고 볼 수 있으며 이를 바탕으로 향후 지속적으로 생성되는 페이스북 게시글들에 대한 목적변수를 할당하여 보다 강건한 스팸 탐지 모델을 구축할 수 있을 것이다.

Table 11. Test Data Confusion Matrix

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

6. 결 론

본 연구에서는 최근 스팸정보의 유통경로가 기존 E-mail과 휴대폰 문자메시지에서 SNS로 변화하고 있는 점과 해당 한글 연구가 거의 없다는 점을 고려하여 대표적인 SNS인 페이스북 게시글들에 대한 스팸 필터링 시스템을 제안하였다. 페이스북 게시글의 경우 명시적인 목표변수가 없는 상황과 실제 스팸 데이터가 비교적 소량으로 존재하여 데이터 수집이 힘든점을 보완하기 위해 인스타그램의 해시태그 정보를 이용하여 스팸성 게시글과 정상 게시글을 수집하여 텍스트 기반 스팸 필터링 시스템을 구축하였다. 또한, 인스타그램 기반의 데이터를 기반으로 한 전이학습 결과와 페이스북 게시글의 메타 정보를 활용한 변수를 추가함으로써 보다 정확한 스팸 탐지 모델을 구축하였다. 인스타그램의 150개의 스팸/햄 해시태그 기준 761,861개의 게시글을 통해 주요 단어를 추출하여 총 11개의 페이스북 게시글 메타정보와 결합하여 랜덤포레스트 기반의 분류기를 학습한 결과, 테스트 데이터 기준 정분류율 99.9%, F1-measure 98.2%의 탐지 성능을 나타내는 것을 확인하였다.

본 연구는 실제 범주의 표식 정보가 부재한 상황에서 페이스북 스팸 필터링 알고리즘을 학습하기 위해 인스타그램의 데이터를 활용하여 전이학습을 행하는 것이 매우 효과적이라는 것을 입증했다는 것에 의의가 있다. 페이스북의 스팸/정상 게시글의 표식 정보를 확보하기 위해 인스타그램의 해시태그를 활용한 결과 두 범주 모두 균형적으로 데이터를 확보 할 수 있었으며, 페이스북의 스팸의 특징을 반영한 메타변수를 통하여 높은 정확도의 스팸 필터링 시스템 구축이 가능하게 되었다. 본 연구에서 제안한 방식은 실제 SNS와 비슷한 다양한 인터넷 환경에 효과적으로 적용될 수 있을 것으로 기대한다.

본 연구에서는 페이스북 전체 페이지 기반 예측을 목적으로 분류기를 생성하였지만 특정 주제로 이루어진 하나의 페이스북 페이지는 대부분의 비슷한 텍스트로 이루어져 있을 확률이 있으므로, 개별 페이지별로 콘텐츠의 주제를 분석하는 토픽 모델링 기법 등을 기반으로 스팸 필터링 시스템을 구축해볼 수 있을 것이다. 또한 해당 연구 방법론을 통하여 지속적으로 페이스북 데이터를 습득한다면, 인스타그램의 학습된 단어가 아닌 페이스북의 대량의 문서(corpus)데이터를 통한 분류기를 생성 할 수 있을 것이다. 또한 스팸 게시글의 주제는 지속적으로 진화하는 특징을 가지고 있기 때문에, 새롭게 생성되는 스팸 주제에 대한 데이터 습득을 어떻게 습득할 것인지도 향후 연구되어야 할 것이다. 또한, 본 연구에서는 페이스북 API 정책상 사용자에 대한 정보를 습득하는 것에 한계가 있어 사용자에 대한 정보를 통해 스패머를 추출하는 알고리듬을 같이 사용하지 않았지만 해당 데이터를 가용 할 수 있다면 본 연구를 통해 제시한 분류기와 함께 사용하여 향상된 스팸 필터링 시스템 구축이 가능하게 될 것으로 기대한다.

참고문헌

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

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

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

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

Joe, I. H. and Shim, H. T. (2009), A SVM-based Spam Filtering System for Short Message Service, *The Korean Institute of Commu-*

Table 12. Spam Filtering Performance of the Proposed Model for the Holdout Dataset

Valid index / Proposed method	Sensitivity	Precision	Specificity	Accuracy	BCR	F1
Predict results of test data	1.000	1.000	0.964	0.999	0.982	0.982

nications and Information Sciences, **34**(9), 908-913.

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

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

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

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

Oh, Y. H., Kim, H., Yoon, J. S., and Lee, J. S. (2014), Using Data Mining Techniques to Predict Win-Loss in Korean Professional Baseball Games, *Journal of Korean Institute of Industrial Engineers*, **40**(1), 8-17.

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.

Shannon, C. E. (2001), A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review*, **5**(1), 3-55.

Soiraya, M., Thanalerdmongkol, S., and Chantrapornchai, C. (2012), Using a Data Mining Approach : Spam Detection on Facebook, *International Journal of Computer Applications*, **58**(13), 26-31.

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

Yang, C., Harkreader, R. C., and 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.

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

Zhang, X., Li, Z., Zhu, S., and Liang, W. (2016). Detecting spam and promoting campaigns in Twitter, *ACM Transactions on the Web (TWEB)*, **10**(1), 4:1-28.

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