앙상블 머신러닝 모델 기반 유튜브 스팸 댓글 탐지

Ensemble Machine Learning Model Based YouTube Spam Comment Detection

저자
(Authors)

정민철, 이지현, 오하영
Min Chul Jeong, Jihyeon Lee, Hayoung Oh

출처
(Source)

한국정보통신학회논문지 24(5), 2020.5, 576-583(8 pages)
Journal of the Korea Institute of Information and Communication Engineering 24(5), 2020.5, 576-583(8 pages)

발행처
(Publisher)

한국정보통신학회
The Korea Institute of Information and Communication Engineering

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09349306

APA Style

정민철, 이지현, 오하영 (2020). 앙상블 머신러닝 모델 기반 유튜브 스팸 댓글 탐지. 한국정보통신학회논문지, 24(5), 576-583

JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 5: 576~583, May. 2020

앙상블 머신러닝 모델 기반 유튜브 스팸 댓글 탐지

정민철[1] · 이지현[2] · 오하영[3*]

Ensemble Machine Learning Model Based YouTube Spam Comment Detection

Min Chul Jeong[1] · Jihyeon Lee[2] · Hayoung Oh[3*]

[1]Undergraduate Student, Department of Digital Media, Ajou University, Suwon, 16499 Korea
[2]Undergraduate Student, Department of English Language and Literature, Ajou University, Suwon, 16499 Korea
[3*]Assistant Professor, Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

요 약

이 논문은 최근 엄청난 성장을 하고 있는 유튜브의 댓글 중 스팸 댓글을 판별하는 기법을 제안한다. 유튜브에서는 광고를 통한 수익 창출이 가능하기 때문에 인기 동영상에서 자신의 채널이나 동영상을 홍보하거나 영상과 관련 없는 댓글을 남기는 스패머(spammer)들이 나타났다. 유튜브에서는 자체적으로 스팸 댓글을 차단하는 시스템을 운영하고 있지만 여전히 제대로 차단하지 못한 스팸 댓글들이 있다. 따라서, 유튜브 스팸 댓글 판별에 대한 관련 연구들을 살펴보고 인기 동영상인 싸이, 케이티 페리, LMFAO, 에미넴, 샤키라의 뮤직비디오 댓글 데이터에 6가지 머신러닝 기법(의사결정나무, 로지스틱 회귀분석, 베르누이 나이브 베이즈, 랜덤 포레스트, 선형 커널을 이용한 서포트 벡터 머신, 가우시안 커널을 이용한 서포트 벡터 머신)과 이들을 결합한 앙상블 모델로 스팸 탐지 실험을 진행하였다.

ABSTRACT

This paper proposes a technique to determine the spam comments on YouTube, which have recently seen tremendous growth. On YouTube, the spammers appeared to promote their channels or videos in popular videos or leave comments unrelated to the video, as it is possible to monetize through advertising. YouTube is running and operating its own spam blocking system, but still has failed to block them properly and efficiently. Therefore, we examined related studies on YouTube spam comment screening and conducted classification experiments with six different machine learning techniques (Decision tree, Logistic regression, Bernoulli Naive Bayes, Random Forest, Support vector machine with linear kernel, Support vector machine with Gaussian kernel) and ensemble model combining these techniques in the comment data from popular music videos - Psy, Katy Perry, LMFAO, Eminem and Shakira.

키워드 : 데이터 분석, 분류, 스팸 댓글, 앙상블 머신러닝, 유튜브 댓글

Keywords : Data analysis, Classification, Spam Comment, Ensemble Machine Learning, Youtube comment

Received 13 November 2019, **Revised** 17 November 2019, **Accepted** 19 November 2019

*** Corresponding Author** Hayoung Oh (E-mail: hyoh79@gmail.com Tel:+82-2-583-8585)
Assistant Professor, Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.5.576 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

유튜브는 2005년 창립되고 2006년에 구글에 인수된 세계 최대의 동영상 공유 사이트로, 최근 온라인 콘텐츠의 흐름이 동영상으로 바뀌면서 유튜브가 동영상 콘텐츠 플랫폼으로써 엄청난 성장을 하고 있다. 현재 유튜브에 1분 동안 업로드 되는 동영상의 분량은 400시간 이상이고, 전 세계에서 1분 동안 시청하는 동영상의 수도 450만 건 이상이다[1]. 동영상 시청뿐만 아니라 동영상 업로드까지 쉽게 할 수 있는 유튜브의 접근성으로 인해, 1인 미디어가 많아졌고 몇몇 일반인들은 온라인에서 유명해지면서 사회에 영향을 끼치는 인플루언서가 되었다.

유튜브 크리에이터들은 구독자의 수가 1,000명 이상이고 최근 12개월간 채널의 시청 시간이 4,000시간 이상이라면 동영상 조회수와 시청 시간에 따른 광고 노출 정도로 수익 창출이 가능하다[2]. 이에 따라, 그림 1과 같이, 인기 동영상에 자신의 채널이나 동영상을 홍보하는 스팸 댓글들이 작성되고 있다. 또한, 동영상과 관계없는 정치적 댓글이나 욕설, 비하 발언 등의 악플로 인해 댓글 기능을 닫아놓은 크리에이터들도 있다. 따라서, 유튜브에서는 자체적인 스팸 차단 시스템을 운영하고 있지만, 여전히 잡아내지 못한 스팸 댓글들이 있다. 기존 유튜브 스팸 댓글 탐지에 대한 연구들에서는 여러 가지 머신러닝 기법들을 각각 데이터 셋에 적용하여 댓글을 분류하고 성능을 비교하였다. 따라서, 본 논문에서는 여러 모델의 결과를 결합하여 최종 결과를 도출하는 앙상블 머신러닝 기법을 제안한다.

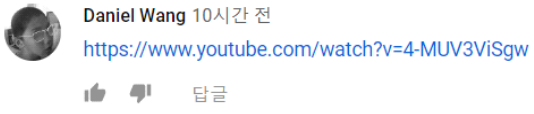

Fig. 1 Examples of Spam comment for promoting self video

논문은 구성은 다음과 같다. 2장에서는 관련 연구를 서술한다. 3장에서는 시스템 모델 및 제안하는 기법에 관해서 설명하고 4장에서는 실험 및 결과를 설명한다. 마지막 5장에서는 결론을 맺는다.

Ⅱ. 관련 연구

기존 스팸 관련 연구들은 SMS 문자 서비스, 이메일, 포탈 사이트 및 블로그등과 같은 웹사이트를 대상으로 이루어졌다. 하지만, 최근 스팸 연구 동향을 살펴보면 유튜브 스팸에 관한 연구가 활발히 이루어지고 있음을 확인할 수 있다. 그 이유는 다음과 같다. 유튜브가 등장하고 그 규모가 상당해지면서 유튜브는 가장 영향력 있는 소셜 커뮤니티 역할을 하게 되었다. 유튜브는 동영상 공유 사이트로 입지를 강화하고 있으며 사용자와 콘텐츠의 수 또한 지속적으로 증가시키고 있다. 하지만 이는 홍보와 금전적 목적을 달성하기에 용이한 환경을 조성해서 각종 불법 스팸 영상과 댓글이 증가하게 되었다.

따라서, 본 연구에서는 스패머 인식 및 탐지와 관련된 기존 연구들을 웹사이트 상의 스팸과 유튜브의 스팸 두 범주로 나누어 살펴보고자 한다.

2.1. 웹사이트에서의 스팸 탐지

소비자들은 상품 구매 후 웹사이트에 후기를 올리는데, 단어를 분해하는 방법 중 하나인 n-gram 모델을 사용해서 이러한 후기의 규칙성을 찾아내는 연구가 있다[3]. 즉, n-gram은 사전을 사용하는 데는 어느 정도 한계가 있고, 컴퓨터가 자동으로 품사를 분해하는데도 규칙에 따르지 않는 경우도 존재하기 때문에 차라리 적당한 크기로 잘게 쪼개서 분석하자는 목적을 가지고 있다.

웹 사이트에서 스팸 및 악성 후기는 복제되었다는 가정 하에 세 유형으로 나뉘는데, 그 중 같은 내용을 다른 아이디로 업로드 한 경우에 대해 이 모델을 사용했다. 즉, 상품 후기 데이터의 언어 분석, 특히 특정 단어 뒤에 어떤 단어가 위치하는지 예측하면서 각각을 비교하고 무분별하게 복제된 후기 데이터를 찾아내기 위해 n-gram 언어 모델을 사용했다.

블로그 스팸을 다룬 연구[4][5] 또한 언어 모델 중 하나를 사용했다. 해당 연구에서는 데이터 수집을 위해 랜덤으로 블로그 포스팅 50개와 해당 포스팅에 달린 1024개의 댓글을 크롤링 했다. 댓글 데이터는 휴리스틱하게 직접 일반 댓글과 스팸 댓글로 분류되었고, 이 중 32%는 일반 댓글 그리고 68%는 스팸이었다. 블로그 포스팅, 댓글과 댓글에 포함된 외부 링크의 글쓰기 방식이 다르기 때문에, 각각에 언어 모델을 적용하였다. 이 모델들의 유사도를 통해 댓글의 스팸 여부를 확인하였다. 즉,

언어 모델은 어떤 단어가 뒤에 오는지에 대한 예측을 바탕으로 단어 순서에 확률을 할당하는 모델이기 때문에 이를 활용한다면 스팸 탐지의 주요한 방법이다.

2.2. 유튜브에서의 스팸 탐지

유튜브 댓글은 위 연구들과 달리 언어 모델을 적용하기에 많은 어려움이 따른다. 유튜브 영상에 달린 댓글은 길이가 짧고 그 자체로의 정보가 부족하기 때문에 영상과의 관련성이 비교적 적기 때문이다. 따라서 유튜브 스팸 댓글은 기존 스팸 인식 연구들과 달리 분류 알고리즘에 초점을 두고 있다. 또한, 유튜브라는 플랫폼의 특성상 스팸은 영상과 댓글의 형태로 각각 나타날 수 있기 때문에 유튜브 내에서도 스팸의 종류가 나누어져 있다. 이를 토대로 영상 스팸을 다룬 연구를 먼저 살펴보고 분류 기법을 적용한 댓글 스팸 관련 연구를 소개할 것이다.

현재 유튜브에서는 없어진 기능이지만 특정 영상에 대한 반응으로 댓글 외에 영상을 게시할 수 있었다. [6]의 저자들은 답글 영상에서 스팸을 찾아내고자 실험 기간 동안에 답글 영상을 주고받은 사용자와 해당 영상을 랜덤으로 수집했다. 592명의 유튜브 계정을 대상으로 16,611명의 스팸 사용자들을 찾았고, 이들이 총 8,710개의 영상에 스팸 답글 영상을 달았다는 것을 확인하였다. 스팸 답글 영상은 답글 영상이 목표로 하는 영상의 주제와 관련 없이 특정 상품을 홍보하거나 외설적 내용을 다룰 경우 스팸으로 분류됐다. 이렇게 분류된 데이터에 Support Vector Machine (SVM) 분류 기법을 적용하고 스팸 매트릭스, 정확도와 F1 스코어로 평가했다. 해당 기법에서 SVM 분류 기법은 입력된 데이터가 스팸에 속할 확률이 얼마나 되는지 지속적으로 학습 및 계산 후 업데이트되면서 분류되는 기법이다. 이를 토대로 유튜브 사용자와 영상의 특성을 파악하고 답글 영상이라는 유튜브 내 사회적 연결망을 확인하였다.

[7]의 연구에서는 유튜브 영상의 특징을 추출하여 스팸 여부를 가리고 분류 알고리즘을 적용하였다. 데이터 수집 기간 동안 영상을 올린 500명의 채널을 찾고 영상 30,621개를 크롤링했다. '채널 생성 일자', '업로드 된 영상의 평균 수', '채널 생성일에 따른 조회수 비율'과'전체 조회수에 대한 좋아요 비율' 등 채널의 특징 16개로 한정하여 각 채널을 분석하였다. 그런 다음 9개의 알고리즘을 적용해서 해당 특징이 스팸 탐지에 얼마나 기여하는지 평가하였고, 결과 Bayes Network, Naïve Bayesian과 Bayes classifier가 약 98%의 정확도를 보여주었다.

유튜브 스팸 탐지에도 N-gram 모델을 사용한 최신 연구[8]가 있다. 유튜브 API로 최근 업로드된 영상과 댓글을 불러와서 Multinomial Naïve Bayes, Random Forests 그리고 SVM 분류 기법을 적용한 연구다. F1 스코어로 각각을 평가한 결과, SVM는 0.9774% 그리고 Random Forests는 0.9726%의 높은 정확도를 도출했다.

[9]의 연구는 가장 많이 본 영상 10개 중 5개 영상에 대한 댓글을 수집해서 이를 'Labeling'이라는 협업 태그 도구를 통해 직접 스팸과 일반 댓글로 분류하였다. 데이터 전처리 과정에서는 Bag of words (BOW)[10]와 Term frequency (TF)가 사용되었다. Bag of words란 단어의 빈도수만 고려하여 문서를 표시하는 방법으로 단어의 순서 무시, 단어의 빈도를 행렬로 표현하는 TDM에 기반을 둔다. 예를 들어, 여러 개의 문장들의 Bag of words 및 TDM에서 각 단어들의 빈도수는 유사하지만 해당 텍스트들을 직접 비교해서 읽어보면 전혀 다른 뜻을 가진 텍스트라는 경우가 비일비재하기 때문에 단순히 빈도수에 의해 키워드를 파악하는 BOW 방식의 문장 표현은 한계점이 있다. 그럼에도 불구하고 [9]연구에서는 데이터 전처리과정에서 이를 활용했다.

70%는 훈련 데이터로, 30%는 테스트 데이터로 사용되고 새로운 댓글 데이터는 알고리즘에서 테스트 할 수 있도록 추가되었다. 각 알고리즘의 성능을 비교하기 위해 정확도, 스팸 탐지 비율, 정상 댓글 차단 비율과 F1 스코어가 사용되었다. 이와 더불어 MCC를 사용하여 정상 댓글이 차단된 경우의 비율도 같이 고려했다. 분류 기법은 10개(CART, K-NN, LR, NB-B, NB-G, NB-M, RF, SVM-L, SVM-P, SVM-R)가 사용되었다. 각각을 적용한 결과, 연구에서 사용된 10개 기법 모두 90%이상의 정확도와 일반 댓글 차단 비율은 5%에 불과했다. 그 중 CART, LR, NB-B, RF, SVM-L and SVM-R는 99.9%의 신뢰성을 보여주었다.

[11]은 [9]의 연구를 보완하고자, [9]에서 사용하지 않은 ANN 모델을 사용하여 같은 데이터 셋으로 더 나은 스팸 탐지 결과를 제시했다. ANN 적용 후, 정확도, 스팸 탐지 비율, 정상 댓글 차단 비율, F1 스코어와 MCC를 기준으로 [9]과의 결과를 비교하였다. 그 결과 ANN이 다른 모델보다 높은 정확도, F1 스코어와 MCC를 보였다.

기존 Rafaqat의 연구[12]에서는 유튜브 댓글의 스팸

탐지에 관한 연구에 대해 데이터 셋과 사용한 기법을 중심으로 결과를 비교 및 정리하였다. 제시된 선행 연구는 최소 2개 이상의 분류 기법을 사용했고 각각 높은 정확도를 보였다. 좋은 결과를 내는 기법이 데이터 셋마다 다르기 때문에, 여러 기법으로 실험을 하며 최상의 결과를 내는 분류 알고리즘을 찾아야 됨을 제시했다.

Ⅲ. 시스템 모델 및 제안하는 기법

3.1. 실험 방법 및 실험 환경

제안하는 기법은 유튜브 스팸 댓글 탐지에 대한 대표적인 선행 연구인 Tulio의 연구[9]를 비교 연구 대상으로 삼았다. 이를 위해, [9]에서 좋은 성능을 보였고, 99.9%의 신뢰수준에서 유의미했던 6가지 머신러닝 기법 CART (의사결정나무, Decision Tree), LR (로지스틱 회귀분석, Logistic Regression), NB-B (베르누이 나이브 베이즈, Bernoulli Naive Bayes), RF (랜덤 포레스트, Random Forest), SVM-L (선형 커널을 이용한 서포트 벡터 머신, Support vector machine with linear kernel), SVM-R (가우시안 커널을 이용한 서포트 벡터 머신, Support vector machine with Gaussian kernel)들과 이를 결합한 앙상블 모델을 제안하고 성능평가를 진행했다.

실험 환경은 주피터 노트북에서 파이썬 3.7.1버전과 사이킷런 라이브러리 0.20.1버전을 사용하였다[13,14,15].

3.2. 제안하는 기법의 실험 개요

제안하는 기법의 실험 개요는 그림 2와 같다. 댓글 데이터 1983개를 수집하고 전체의 70%인 1369개를 훈련 데이터, 30%인 587개를 시험 데이터로 나눈다. 전처리를 한 후에 훈련 데이터를 머신러닝 알고리즘 6가지 기법(CART, LR, NB-B, RF, SVM-L, SVM-R)과 이들을 결합하여 만든 ESM-H (Ensemble with hard voting), ESM-S (Ensemble with soft voting)으로 학습을 시키고 분류 모델을 만든다. 만들어진 모델에 시험 데이터를 넣어서 클래스를 예측하고 평가한다.

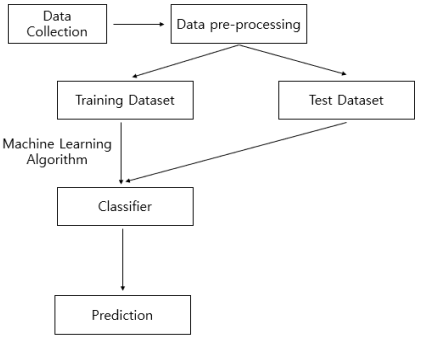

Fig. 2 Overview of the proposed spam comment detection scheme

3.3. 데이터 셋

본 논문에서는 [9]에서 제공한 5가지 인기 뮤직비디오의 댓글 데이터로 실험을 진행하였다. 데이터는 [16]에서 다운로드가 가능하다. 데이터는 댓글 ID, 댓글 작성자, 날짜, 댓글 내용, 레이블 링 된 클래스(0: Ham 혹은 1: Spam)로 이루어졌고, 본 논문에서는 댓글 내용과 레이블 링 된 클래스만을 사용하였다. 5가지 데이터를 각각 훈련(train)시키고 시험(test)하면 5개의 분류기가 해당 데이터에 대해서만 성능이 좋고 다른 동영상의 댓글 데이터에는 적용이 잘되지 않는 과적합(overfitting)이 발생할 수 있다. 따라서, 본 논문에서는 표1과 같이 일반화를 위하여 5가지 데이터를 모두 결합한 1983개의 데이터 중에 70%인 1369개의 데이터를 훈련 데이터, 30%인 587개의 데이터를 시험 데이터로 분리하여 실험하였다.

Table. 1 Datasets collected and used in the experiments

Datasets	Spam	Ham	Total
Psy	175	175	350
KatyPerry	175	202	350
LMFAO	236	202	438
Eminem	245	203	448
Shakira	174	196	470
Total	1005	978	1983

3.4. 데이터 전처리

댓글 데이터는 텍스트 데이터이기 때문에 머신러닝 기법에 적용할 수 있도록 전처리를 하였다. 전처리는 불용어(stopwords) 제거를 하고 파이썬 사이킷런 라이브러리의 CountVectorizer 함수를 사용하여 댓글을 토큰의 목록으로 만들고 토큰의 출현 빈도를 카운트한 후에 Bag of words(BoW) 벡터화를 하였다. 해당 과정은 그림 3과 같다.

Comment : “John likes to watch movies. Mary likes movies too.”

1. tokenize
[“John”, “likes”, “to”, “watch”, “movies”, “Mary”, “likes”, “movies”, “too”]

2. Bag of words
BoW = {“John”:1, “likes”:2, “to”:1, “watch”:1, “movies”:2, “Mary”:1, “too”:1}

Fig. 3 Tokenize and BoW vectorize process[17]

3.5. 머신러닝 알고리즘

3.5.1. 의사결정나무 (Decision Tree)

의사결정나무는 머신러닝 기법 중에 분류나 예측을 위한 방법으로, 나무 구조를 사용해서 전체 데이터를 소집단으로 분리한다. 부모노드(Parent node)에서 자식노드(Child node)로 나뉘는 것을 분기(Splitting)라고 하며 분기에 사용되는 변수와 기준 값에 따라 나무의 구조가 달라진다. 변수와 기준 값은 부모노드와 이 기준에 의해 분기되는 자식노드들의 지니 불순도(Gini impurity) 혹은 엔트로피(Entropy)를 계산하여 정보이득(Information gain)이 큰 변수와 기준 값을 선택한다. 가장 상위 노드를 뿌리노드(Root node)라고 하며, 뿌리노드에는 분류 대상이 되는 모든 데이터가 포함되고 분기를 할 때마다 집단이 분리된다. 최종적으로 의사결정나무의 최대 깊이를 조절하는 등의 가지치기(Pruning)를 하여 모델을 결정한다.

3.5.2. 로지스틱 회귀분석 (Logistic Regression)

로지스틱 회귀분석(Logistic regression analysis)은 다양한 분야에 매우 널리 사용되는 분석방법으로 결과가 몇 가지 범주로 나눠지는 경우 유용하게 사용할 수 있다. 특히, 로지스틱 회귀분석은 종속변수가 이진 반응변수로 분류되는 경우에 주로 사용될 수 있다.

예를 들어, 비만에 의한 당뇨 발생률의 차이 등의 위험 요인에 따른 질병 발생의 유무 혹은 수술 후 생존에 대한 반응 변수가 0과 1사이의 확률 값으로 나타내고 사망일지 생존일지를 통계적으로 분석하고 할 때 사용될 수 있다.

3.5.3. 나이브 베이즈 (Naive Bayes)

나이브 베이즈는 분류 문제에 베이즈 이론을 적용한 지도학습 알고리즘으로 텍스트 분류에 주로 많이 이용된다. 분류기가 실행되기 전에 훈련 데이터를 활용해 특징 값이 제공하는 증거를 기반으로 결과가 관측될 확률을 계산하는 학습 벡터를 통한 학습이 이뤄져야하며 분류자는 레이블이 없는 새로운 데이터에 대해서 각 계층에 속할 확률을 추정하고 가장 높은 확률을 가진 계층으로 예측하여 객체를 분류한다.

나이브 베이즈의 장점으로는 간단하고 빠르고 증거를 나타내는 특징 벡터를 계산에 모두 포함시킬 수 있기 때문에 저장 공간과 계산 시간 측면에서 매우 효율적이며 잡음과 누락 데이터를 잘 처리 한다는 것이다. 또한, 훈련에는 사례를 조사하면서 단지 계층과 특징이 나타내는 횟수만 저장하면 되기 때문에 상대적으로 적은 예시가 필요하지만, 대용량의 예시에도 매우 잘 작동 된다. 특히, 예측을 위한 추정 확률을 쉽게 얻을 수 있다는 특징이 있기 때문에 실무에서 뛰어난 성능으로 인지도가 높다. 하지만, 데이터 셋의 모든 특징이 동등하게 중요하고 독립적이라고 가정하기 때문에 때로는 실제 응용에는 맞지 않는 경우도 존재한다.

3.5.4. 서포트 벡터 머신 (Support Vector Machine)

서포트 벡터 머신은 다차원 공간에 표시되는 점들 사이에 초평면이라는 경계를 만드는 표면 경계를 사용해 데이터를 유사한 클래스 값들의 그룹으로 분할하는 기법이다. 즉, SVM의 목표는 공간을 나눠 양쪽에 매우 균질적인 분할을 생성하는 초평면이라고 하는 평평한 경계를 생성하는 것이다. SVM은 선형이나 비선형 분류, 회귀 등 다양한 분야에서 사용이 가능하기에 다목적 머신러닝 모델이며, 머신러닝 알고리즘 중에서 가장 유명한 알고리즘 중 하나로 적당한 데이터 사이즈를 가지고도 아주 좋은 효과를 내기 때문에 활용도가 높다.

차원별 SVM의 작동 원리는 다음과 같다. 2차원에서 SVM 알고리즘의 작업은 두 클래스 사이를 가장 멀리 분리하는 여러 분할 직선들 중에서 최대 마진(Margin) 초평면 직선을 찾는 것이다(Separating Hyperplane). 가장 멀리 분리하게 만드는 직선이 미래 데이터에 대해 가장 일반화를 잘할 것이며 최대 마진은 임의의 잡음이 있더라도 점이 경계의 올바른 쪽에 남게 될 가능성을 높이기 때문이다. 즉, 서포트 벡터(Support Vectors)의 정확한 정의는 각 클래스에서 최대 마진 초평면 MMH에 가장 가까운 점들을 의미한다. SVM은 특징의 개수가 엄청나게 많더라도 분류 모델을 저장하기 위한 아주 간결한 방법을 제공하며, 신경망 사용보다 쉬운데, 특히 잘 지원되는 SVM 알고리즘들이 있기 때문이다

반면, SVM은 최고의 모델을 찾기 위해 커널과 모델 파라미터의 다양한 조합을 테스트해야 하며, 훈련이 느릴 수 있으며, 해석하기 어려운 복잡한 블랙박스 모델이 만들어질 수 있다. 마지막으로 선형적으로 분리할 수 없는 경우에는, 비선형 공간을 위한 커널(kernel)을 사용해서, 분리할 수 있는 수준으로 차원을 올려줘야 한다. 즉, 2차원상에서 분류가 어려운 집단 간의 분류는 3차원으로 확대하여 분류할 수 있다.

3.6. 제안하는 기법

3.6.1. 앙상블 (Ensemble) 머신러닝 기반 유튜브 스팸 댓글 탐지 기법

앙상블 모형이란 주어진 데이터를 이용하여 여러 개의 서로 다른 예측 모형을 생성한 후, 이러한 예측 모형의 예측 결과를 종합하여 하나의 최종 예측결과를 도출해 내는 방법이다[18].

본 연구에서는 먼저, 단순 다수결 방식을 사용하여 더 많은 분류기가 선택한 클래스를 최종 클래스로 채택하도록 하는 **ESM-H** (Ensemble with hard voting) 모델을 제안한다. 즉, 5개의 분류기 중에 3개의 분류기가 클래스 0, 2개의 분류기가 클래스 1로 예측을 하였다면 클래스 0이라고 예측하는 앙상블 모델을 활용한다. 따라서, 투입한 분류기의 수를 홀수로 만들기 위해 표. 2와 같이 기존 6가지 기법 중 성능이 가장 낮았던 NB-B를 제외한 나머지 5가지 기법으로 생성하였다. 두 번째로는 각 분류기가 클래스를 예측한 확률들의 평균을 내어 최종 예측을 하는 **ESM-S** (Ensemble with soft voting) 모델을 제안하였다.

Ⅳ. 실험 및 결과

표. 1에 있는 전체 데이터의 70%를 훈련 데이터, 30%를 시험 데이터로 나누고 8가지 머신러닝 기법들을 적용하여 실험을 진행하였다. 사용한 기법들은 표 2와 같다.

성능 측정은 Acc (Accuracy rate, 정확도), SC (Spam caught rate, 스팸 탐지 정도), BH (Blocked ham rate, 정상 댓글 차단 정도), F1-score, MCC (Matthews correlation coefficient, 매튜 상관계수)를 사용하였다. 각각 구하는 공식은 표 3의 confusion matrix를 참조했을 때 다음과 정리된다.

Table. 2 Classification methods used in the experiments

Classification methods	
CART	Decision Tree
LR	Logistic Regression
NB-B	Benoulli Naive Bayes
RF	Random Forest
SVM-L	Support vector machine with linear kernel
SVM-R	Support vector machine with Gaussian kernel
ESM-H	Ensemble with hard voting
ESM-S	Ensemble with soft voting

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$SC = Recall\ \ of\ \ Spam \tag{4}$$

$$BH = 1 - Recall\ \ of\ \ Ham \tag{5}$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recalll} \tag{6}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

Table. 3 Confusion matrix. Positive : Spam, Negative : Ham

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Table. 4 Experiment result

Methods	Acc(%)	SC(%)	BH(%)	F1-score	MCC
CART	93.53	92.18	5.00	0.935	0.871
LR	92.33	88.27	3.21	0.923	0.851
NB-B	85.87	73.62	0.71	0.858	0.747
RF	92.84	89.58	3.37	0.928	0.860
SVM-L	94.21	93.81	5.36	0.942	0.884
SVM-R	93.19	89.58	2.86	0.932	0.867
ESM-H	94.38	92.18	3.21	0.944	0.889
ESM-S	**95.06**	**93.16**	2.86	**0.951**	**0.902**

실험 결과, 표4에서 보여주듯이 Acc, SC, F1-score, MCC에서 ESM-S 모델이 가장 좋은 성능을 보였고 BH에서는 ESM-S 모델이 NB-B 기법 다음으로 좋은 성능을 보였다. 8가지 분류기로 False Positive rate {FP/(FP+TN)}를 x축, Recall을 y축으로 사용하여 만든 ROC (Receiver Operating Characteristic) 최종 커브는 그림 4와 같다. ROC 커브의 밑면적인 AUC (Area Under a ROC Curve)는 면적이 1에 가까울수록 TP(스팸을 스팸이라고 예측한 것)와 FN(정상 댓글을 정상이라고 예측한 것)이 동시에 높은 것이므로 좋은 커브다. 따라서, 회색 곡선으로 나타난 ESM-S모델이 가장 면적이 넓은 것을 확인하였다.

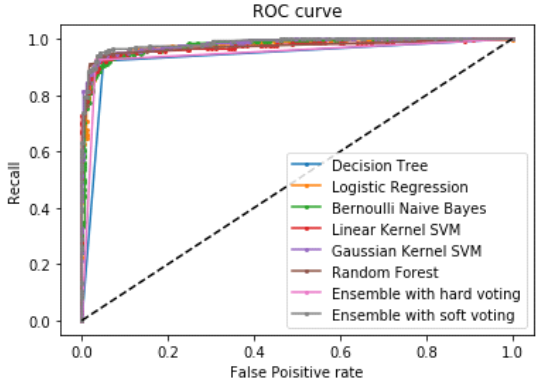

Fig. 4 ROC curve of the proposed classifiers scheme

Ⅴ. 결 론

본 논문에서는 머신러닝 기법들로 앙상블 모델을 만들어 유튜브 스팸 댓글 탐지를 하였다. 실험 결과, 본 논문에서 제안한 ESM-S 모델은 5종류의 성능 평가 중 4종류에서 가장 좋은 성능을 보였고 1종류에서 두 번째로 좋은 성능을 보였다. 따라서, 한 가지 기법만으로 모델을 만들어서 분류를 했던 기존 연구들과는 달리, 본 논문에서는 여러 가지 기법들의 결과를 결합하여 분류를 하는 앙상블 모델을 제안하여 성능 개선을 하였다. 향후 연구에서는 뮤직비디오가 아닌 다른 카테고리의 동영상에도 제안한 모델을 적용해보고, TF-IDF 전처리 및 딥러닝 기법도 추가한다면 더 좋은 성능이 나올 것으로 예상된다.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2017R1D1A1B03035557).

References

[1] KBS NEWS [Internet] Available: https://mn.kbs.co.kr/news/view.do?ncd=4260664

[2] YouTube Help, [Internet] Available: https://support.google.com/youtube/answer/72857?hl=ko

[3] M. S. Patil, and A. M. Bagade, “Online review spam detection using language model and feature selection.” *International Journal of Computer Applications, 59(7), December 2012, 1-4.*

[4] M. Mishne, G. Carmel, D. David, L. Ronny, “Blocking Blog Spam with Language Model Disagreement.”, *ACM Transactions on Multimedia Computing, Communications, and Applications, May, 2005, 1-6.*

[5] T. Bogers and D. B. Van, “Using Language Models for Spam Detection in Social Book marking”, *Proceedings of ECML/PKDD Discovery Challenge Workshop, 2008, 1-12.*

[6] P. S. Kiran, “Detecting spammers in YouTube : A study to find spam content in a video platform”, *IOSR Journal of Engineering (IOSRJEN), 05(07), July 2015, 26-30.*

[7] Y. Yusof and O. H. Sadoon, “Detecting video spammers in youtube social media”, *Proceedings of the 6th International Conference of Computing & Informatics, April 2017, 228-235.*

[8] A. Shreyas, and S. Nisha, “N-Gram Assisted Youtube Spam Comment Detection”, *Procedia Computer Science, 132, Jan 2018, 174-182.*

[9] A. Tulio, L. Johannes and A. Tiago, “TubeSpam: Comment Spam Filtering on YouTube”, *IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Dec 2015, 1-6.*

[10] Bag-of-words model [Internet] Available: https://en.wikipedia.org/wiki/Bag-of-words_model

[11] A. Thulfiqar, and A. Hussein, and Q. Samir, “YouTube spam comments detection using Artificial Neural Network”, *Journal of Engineering and Applied Sciences, 13(22), 2018, 9638-9642.*

[12] A. Rafaqat, “Spammer Detection: A Study of Spam Filter Comments on YouTube Videos”, *Lahore Garrison Education System, May 2019, 1-6.*
[13] Project jupyter [Internet] Available: https://jupyter.org/
[14] Welcome to Python.org [Internet] Available: https://python.org/
[15] Scikit-learn: machine learning in python [Internet] Available: https://scikit-learn.org/stable/
[16] YouTube Spam Collection v.1, [Internet] Available: http://dcomp.sor.ufscar.br/talmeida/youtubespamcollection
[17] YouTube Spam Collection, [Internet] Available: http://www.dt.fee.unicamp.br/~tiago//youtubespamcollection/
[18] Y. J. Jang, H. J. Kim, and H. J. Jo, “Data Mining”, *KNOU PRESS, 2016, 1-200.*

정민철(Minchul Jeong)
아주대학교 디지털미디어학과 재학

※ 관심분야 : 소셜정보망 및 추천 시스템

이지현(Jihyeon Lee)
아주대학교 영어영문학과 재학

※ 관심분야 : 소셜정보망 및 데이터 마이닝

오하영(Hayoung Oh)
이화여자대학교 컴퓨터 공학 석사
서울대학교 컴퓨터 공학 박사
성균관대학교 글로벌융합학부 조교수

※ 관심분야 : 소셜정보망 및 데이터 분석