JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 20, No. 7, pp. 1429-1437, Jul. 2019

인공신경망을 적용한 악성 댓글 분류 모델들의 성능 비교

김 진 우[1] · 조 혜 인[2] · 이 봉 규[3*]
[1]연세대학교 정보대학원 석사과정
[2]연세대학교 정보대학원 석사과정
[3]연세대학교 정보대학원 교수

A Comparison Study on Performance of Malicious Comment Classification Models Applied with Artificial Neural Network

Jin Woo Kim[1] · Hye In Jo[2] · Bong Gyou Lee[3*]
[1,2,3*]Graduate School of Information, Yonsei University, Seoul, South Korea

[요 약]

악성 댓글의 피해는 유명인을 넘어 단순히 댓글을 읽는 개인 SNS 이용자에게도 확대되고 있다. 본 연구에서는 인공신경망을 적용한 악성 댓글 분류 모델을 이용하여 [명사], [명사 + 형용사], [명사 + 형용사 + 동사], [모든 품사]의 4가지 품사 처리 방식과 RNN, LSTM, GRU 3가지 알고리즘을 적용한 총 12개의 모델의 성능을 비교하였다. 본 연구결과를 통해 도출된 AI 알고리즘을 이용한 텍스트 분류 서비스 연구는 악성 댓글 분류 분야뿐만 아니라 다양한 분야에 적용되어 유용하게 사용될 수 있다.

[Abstract]

Victims of malicious comments are increasing from the public figures including celebrities to personal social network users. Furthermore, users who only use comments can be the victim. This study used the malicious comment classification model applied with artificial neural network to compare the performance of 12 models applied 4 preprocessing methods including [noun], [noun+adjective], [noun+adjective+verb], and [all parts of speech] and 3 algorithms of RNN, LSTM, GRU. The study result would be the reference for successful introduction of AI algorithm-based text categorization service.

색인어 : 인공신경망, 순환신경망, 악성 댓글, 분류모델 성능비교

Key word : Artificial Neural Network, Recurrent Neural Network, Malicious Comment, Comparison Performance

http://dx.doi.org/10.9728/dcs.2019.20.7.1429

Received 25 June 2019; **Revised** 10 July 2019
Accepted 25 July 2019

***Corresponding Author; Bong Gyou Lee**

Tel: +82-2-2123-6524
E-mail: bglee@yonsei.ac.kr

Ⅰ. 서 론

SNS의 확산으로 이용자들은 커뮤니티뿐만 아니라 인터넷 뉴스에도 쉽게 댓글을 달 수 있으며, 댓글을 통해 인터넷 이용자들은 더욱 쉽고 빠르게 의견을 교류 할 수 있다. 하지만 개인의 참여가 확대됨에 따라 긍정적인 영향 외에도, 익명성을 악용한 사이버폭력, 악성 댓글, 매크로를 통한 조작된 댓글들의 이슈가 부상하며 댓글문화 형성에 악영향을 주고 있다. 특히 악성 댓글은 유명인을 넘어 개인 SNS 이용자까지 그 피해자가 점점 증가하는 추세이며, 단순히 댓글을 읽는 이용자들 또한 피해자가 될 수 있다. 경찰청 통계에 따르면, 악성 댓글로 발생한 사이버 명예훼손과 모욕사건은 2016년 14,908건, 2017년 13,348건으로 2012년 5,684건에서 4~5년 만에 3배 가까이 증가했다[1]. 또한 대부분의 악성 댓글이 경찰에 신고 되지 않는 점, 악성 댓글을 읽는 이용자 또한 피해자가 될 수 있는 점을 고려한다면 악성 댓글로 피해를 받는 사람들의 수는 훨씬 더 많을 것으로 추정된다.

실제로 악성 댓글에 공격을 받은 사람은 피해자임에도 불구하고 복잡한 신고절차를 거쳐야한다. 본인인증부터 악성 댓글의 주소와 내용, 게시중단 요청 취지, 게시물 중 구체적인 권리침해 표현, 소명 내용을 기입해야 하고, 필요한 경우엔 첨부파일도 보내야 하는 복잡한 절차를 거쳐야 하며 쓰기는 쉽지만 치료가 힘든 악성 댓글은 사회전반에 있어 해결해야할 과제 중 하나로 지목되고 있다. 정보통신망법 제44조의3항에 따르면 포털들은 모니터링 중 명예훼손성 게시글이 발생할 경우 자체적으로 제거할 수 있다. 하지만 네이버, 다음 등 국내 주요 포탈에 달리는 댓글들은 하나의 게시글에 적게는 수십 개에서 많게는 수만 개씩 달리며 현실적으로 댓글들을 모니터링 한다는 것은 불가능하다 볼 수 있다. 미국 공영 라디오(National Public Radio)의 경우 관리자의 모니터링만으로는 대량의 악성 댓글을 관리 할 수 없다 판단하여 결국 2016년 댓글 란을 폐지시킨 전례가 있다. 2018년 국회입법조사처에서 발행한 자료에 따르면, 악성 댓글의 범주는 모욕적내용, 유언비어 등을 외에도 인종·성별·지역 차별 등 지속적으로 확대되고 있으나, 정보통신망법 제44조 7항에 해당하는 불법정보에 명백하게 해당되지 않을 경우 게시자를 법적으로 처벌 할 수 없으며, 해당 정보의 삭제를 포털 사업자에게 강제하기 어렵다고 한다. 입법적 관점에서도 온라인상 혐오 표현에 대한 규제를 어느 범위까지 강화할 것인지에 대한 논의가 필요하다 언급하였으나, 처벌 대상이 되는 혐오 표현의 범주를 구체적으로 제시하고 있지는 못하고 있다. 더욱이 법적 규제를 강화할 경우 자의적인 법률 해석과 과도한 규제로 국민의 정치적인 의사 표현을 제약할 수 있다는 점, 극히 일부 정치적 통제가 심한 국가를 제외하고는 해외에 유사한 법률 사례를 찾기 어려우며, 댓글 규제에 대하여는 입법적으로 검토해야 될 과제가 많이 남아있고, 법적 규제 강화를 통해 해결하기는 힘들다고 목소리가 나오고 있다[2]. 입법적 관점에서 악성 댓글을 해결하기 어려움을 인식하여 기업과 학계에서는 '악성댓글 문제는 오히려 법이 아니라 기술을 통해 메꿀 수 있다'라는 논의와 함께 AI를 이용한 악성 댓글을 분류하는 시스템을 개발하고 있다. 실제로 인스타그램은 AI을 이용한 악성 댓글 자동차단 기능이 있으며, 구글은 게시글의 악성지수를 계산해주는 Perspective을 개발하였다. 하지만 인스타그램은 자동 차단 기준을 명시하지 않았으며, 구글의 Perspective는 알고리즘의 불완전함을 인정하고 댓글 란에 적용하고 있지 않다.

본 연구는 인공신경망을 적용한 분류 모델을 이용하여 [명사], [명사 + 형용사], [명사 + 형용사 + 동사], [모든 품사]의 4가지 품사 처리 방식과 RNN(Recurrent Neural Network), LSTM(Long-Short Term Memory), GRU(Gated Recurrent Unit) 3가지 알고리즘을 적용한 총 12개의 모델의 성능을 비교하고 있다. 본 논문의 구성은 서론에 이어서 2장에서는 적용된 이론 및 배경에 대하여 기술하고, 3장에서 연구설계에 대해 설명하고 있다. 4장에서는 인공신경망을 적용한 모델들의 성능들을 비교 분석하고, 마지막 장에서 연구의 결론 및 한계점에 대해 서술하고 있다.

Ⅱ. 이론적 배경

2-1 온라인상 혐오 표현과 악성댓글 선행연구

우리말샘에서 정의하는 악성 댓글의 정의를 살펴보면 '인터넷의 게시판 따위에 올려진 내용에 대해 악의적인 평가를 하여 쓴 댓글'이라고 한다[3]. '악의적 내용'과 '혐오 표현'에 대한 분류 방안이 논의되고 있지만 어느 정도까지의 내용을 악의적 내용으로 분류하고, 어느 정도까지의 표현을 혐오표현으로 정의하는지에 대한 기준은 아직 명확하게 제시하지 못하고 있다. 학계에서는 '악의적 표현'이라는 단어보다 '혐오표현'이라는 단어로 주로 연구되고 있다. 혐오 표현은 'hate speech'를 번역한 영어로서 과거 'hate'을 '증오', '적의'로 해석하고 'speech' 또한 '언론', '발언' 등으로 번역하여 혼용되어 사용되었지만 최근에는 '혐오표현'이라는 용어로 번역되는 것이 일반화 되는 추세이다[4]. 홍상수(2015)는 "인종, 종교, 성적 지향성, 정치적 지향성, 국적, 민족, 피부색, 성별 등의 속성에 대해서 발화자가 가진 선입견에 근거하여 이를 공격하는 것"이라고 정의하였다[5]. 혐오표현을 담고 있는 악성댓글은 불특정 다수를 넘어, 여성, 남성, 성소수자, 특정 지역 등을 대상으로 점점 다양해지고 있으며, 혐오 표현의 강도 또한 심각해지고 있다. 특히 특정 집단을 대상으로 한 혐오표현은 개인에게만 피해를 주는 것이 아닌, 사회 전체에 나쁜 영향을 준다는 것에 심각성이 있다. 피해자가 되는 개인의 경우 정신적 · 신체적 피해를 주며, 사회적으로는 특정 집단 간의 갈등을 유발하게 되어 사회적 통합을 저해하기도 한다[6]. 인터넷상에서의 혐오표현, 악성댓글 등의 문제가 대두됨

에 따라 학계에서도 혐오표현의 본질, 법제도적 논의, 사회적 영향 등에 관하여 연구가 시행되고 있다. 김영일(2019)은 인터넷뉴스 보편화에 따른 악성 댓글 증가를 주목하여, 영어권과 한국어권의 악성댓글을 LIWC(Linguistic Inquiry and Word Count)와 KLIWC(Korea Linguistic Inquiry and Word Count)를 이용하여 분석하였다. 분석결과 일반댓글에 비해 악성댓글은 긍정적인 정서, 인지적인 과정을 나타내는 단어가 상대적으로 적게 사용되었으며 화, 속어, 3인칭 단수 등의 단어가 많이 사용되는 것을 관찰하였고, 악성댓글을 작성하는 사람들의 경우 분노의 감정을 조절하지 못하고 충분한 생각 없이 댓글을 작성하는 것을 유추하였다[7]. 양혜승(2018)은 네이버 뉴스의 범죄기사를 대상으로 지역혐오 댓글의 비율과 표적 집단 분포, 그리고 혐오댓글의 유형을 분류하였으며, 타 지역에 비해 전라도지역을 대상으로 발생한 조롱하기, 지역명 단순언급 등이 상대적으로 높게 발생한 것을 확인하였다[8]. 전창영(2018)은 방송통신심의위원회의 심의사례를 내용분석 실시하여, 혐오표현의 유형별 발생 비율과 특징, 방송통신심의위원회의 혐오표현에 대한 심의 동향을 파악하였으며, 처벌의 대상이 되는 혐오표현의 유형을 명확하게 제시하는 것의 필요성을 언급하였다[9]. 홍주현(2016)은 SNS에서의 혐오표현이 증가함을 주목하여, 네트워크 분석을 이용하여 피해자와 피해범위를 유형화 하였다[10]. 연구 결과 사회 집단적 피해자의 혐오표현이 가장 많이 언급되었으며, 이용자들 간의 상호작용도 가장 많이 발생함을 확인하였다. 또한 다른 이용자를 인식하는 SNS에서 수집된 데이터의 특성상 폭력적 표현과 폭언은 '일간베스트' 사이트에 비해 적은 것을 확인하였다.

2-2 기계학습을 이용한 악성댓글 분류 관련 선행연구

입법적 관점에서 악성 댓글을 해결하기 어려움을 인식하여 구글과 인스타그램 등 글로벌 IT 기업에서는 AI를 이용하여 악성 댓글을 분류하는 시스템 개발을 시도하고 있다. 학계에서도 악성 댓글이 사회적 이슈로 부상하는 만큼 기계학습을 이용하여 악성 댓글을 분류하고자 하는 연구가 진행되었다.

LEE(2018)는 뉴스, 온라인 커뮤니티, Twitter의 댓글을 대상으로 악성댓글 분류를 시도하였다. 연구에서 사용된 악성댓글은 연구자 3명 중 2명 이상이 동의할 경우 악성댓글로 조작적 정의 하였으며, 분석결과 뉴스 89.08%, 온라인 커뮤니티 89.97%, Twitter 92.09%의 정확도의 성능을 확인하였다[11]. Aiyar(2018)는 인기 있는 Youtube 채널이나 SNS에 사람이 몰리는 현상을 악용하여 자동화된 스팸 텍스트를 개제하는 사용자들의 증가에 주목하여 기계학습을 이용한 스팸 텍스트 자동분류를 하였고, RF, SVM, MNB 각각 알고리즘의 정확도를 평가한 결과 SVM(98.41%), RF(98.13%), MNB(95.21%) 순으로 나타났다[12]. 김묘실(2006)은 SVM을 이용하여 악성 댓글을 판별하는 시스템을 구현하였으며, 자질(명사, 형용사, 동사) 선택에 따라 성능에 차이가 있음을 검증하였다[13]. 홍진주(2016)는 인터넷 댓글에 특화된 감성사전을 구축하였고, 구축된 감성사전을 이용해 댓글의 감성을 분석하였으며, 이를 SVM을 이용하여 악성 댓글 탐지 가능성을 살펴보았다[14].

학계에서 기계학습으로 악성 댓글을 분류하고자하는 연구가 이루어짐에도 불구하고 악성댓글의 조작적 정의가 부족하거나 누락된 상태로 진행부분이 있다. 기준 없이 악성 댓글을 분류할 경우 그 의미가 퇴색될 우려가 있다. 악성댓글의 악의적이라는 추상적인 개념을 구체화하는데 어려움이 있겠지만 조작적 정의를 제시한 후 기계학습을 시도할 필요가 있다.

2-3 신경망 모델(Neural Networks model)

신경망(Neural Networks)은 인간의 뇌를 본 따서 만든 모델로 여러 개의 뉴런들을 상호 보완적으로 연결하는 것을 기본 작동원리로 하며 입력 값에 대해서 최적의 출력 값을 출력 및 예측한다. 여러 개의 뉴런과 같이 각 노드들은 입력층(Input Layer), 은닉층(Hidden Layer), 출력층(Output Layer)으로 구분된다. 각각의 원형 노드 가운데 가장 앞쪽에 위치한 입력층에 값이 주어지면 은닉층, 출력층으로 전달된다[15]. 신경망들은 입력을 받고, 그 입력에 따라 내부 상태를 변경하며, 입력 및 활성화에 따라 출력을 생성한다.

1) 순환신경망(Recurrent Neural Network, RNN)

순환신경망(Recurrent Neural Network, RNN)은 자연어와 같이 순차적인 정보를 처리하는 데에 특화된 네트워크 모델이다[16]. 실제 신경망은 양방향으로 구성되어 있지만, 과거 기존의 대부분의 인공신경망 구조는 해석하기에 편리함을 위해 신호의 흐름이 입력에서 출력으로 한 방향으로만 전개되는 것이 대부분이었다. 반면 RNN은 신호가 한쪽 방향으로 흘러가는 것이 아닌 순환 구조를 갖는 인공신경망으로 출력된 결과는 이전의 계산 결과에 영향을 받는다. RNN은 자연어 연구나 번역 분야에 적합한 구조를 갖고 있으며, 최근에는 영상 처리나 분석에도 활용이 많이 되고 있다. 고전적인 신경망 모델에서는 모든 입력과 출력이 각각 독립으로 가정하고 있지만 자연어와 같은 실제 데이터는 시간의 흐름에 따라 순차적으로 정보들이 배열되어있으며, 데이터 간에 관계를 갖고 있다. 예를 들어 어떤 문장이 주어졌을 때 문장 내의 단어의 의미는 단순히 단어의 사전적인 의미가 아닌 전후 맥락에 의해 뜻을 해석해야 경우가 많으면 이러한 이유로 많은 자연어 처리 문제를 해결하기위해 RNN이 이용되곤 한다. RNN구조는 아래와 같으며 이를 도식화한 것은 <그림 1>과 같다[17].

$$h_t \equiv \sigma_h(U_h x_t + V_h h_{t-1} + b_h) \quad (1)$$

$$y_t = \sigma_y(W_y h_t + b_h) \quad (2)$$

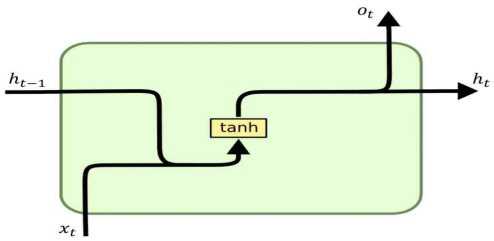

그림 1. Recurrent Neural Network 도식화
Fig. 1. Recurrent Neural Network Structure

2) LSTM(Long-Short Term Memory)

LSTM은 Hochreiter&Schmidhuber(1997)에 의해 처음 소개 되었다. RNN의 한 종류인 LSTM은 RNN구조에 비해 더 긴 시퀀스를 효과적으로 잘 기억하며 다양한 분야에 적용되는 모델 중 하나이다. 순환신경망은 순차적인 데이터가 길어질 경우 이전의 결과가 축적되고, 이로 인해 먼 거리에 있는 상태(state)가 현재의 결과에 미치는 영향이 미미해진다는 단점이 있다. 이를 개선하기 위한 알고리즘이 LSTM이다. LSTM은 RNN에 비해 본질적으로 다른 구조를 갖고 있다고 하긴 힘들지만, hidden state를 계산하는데 다른 식을 사용한다. LSTM의 경우 장기간 메모리 역할을 수행하는 cell state와 연결의 강도를 조절하는 3개의 gate(forget, input, output)로 구성되어 있다. LSTM의 신경망은 gate 조절을 통해 이전 신경망 정보가 현재 신경망에 끼치는 영향을 조절할 수 있으며, 현재 입력과 연관된 정보를 추가할 수도 있고, 다시 출력에 끼치는 영향의 수준을 정할 수 있게 된다. 결과적으로 RNN이 갖고 있던 장기간 메모리가 필요한 문제가 해결이 가능하게 되었으며 많은 분야에서 활용이 되고 있으며, LSTM구조는 아래와 같으며 이를 도식화한 것은 <그림 2>과 같다[17].

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \quad (4)$$
$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \quad (5)$$
$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \quad (6)$$
$$\widetilde{C_t} = \tanh(W_c \bullet [h_{t-1}, x_t] + b_c) \quad (7)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \quad (8)$$
$$h_t = o_t \odot \tanh(C_t) \quad (9)$$

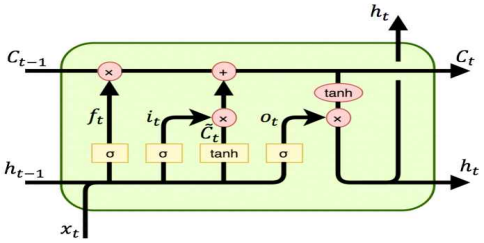

그림 2. Long-short Term Memory 도식화
Fig. 2. Long-Short Term Memory structure

3) GRU(Gated Recurrent Unit)

GRU는 뉴욕대의 조경현 교수가 2014년에 발표한 논문 "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation"에서 처음 소개 되었다. GRU는 LSTM과 마찬가지로 gate를 이용하여 정보의 양을 조절하는 것은 동일하지만, gate의 제어 방식에 있어 차이가 있다. LSTM이 3개의 gate를 이용하는 것에 비해 GRU는 update와 reset 2개의 gate를 보유하고 있다. GRU에서는 LSTM의 forget과 input gate를 결합하여 1개의 update gate를 만들었으며, 별도의 cell state와 hidden state를 hidden state로 묶은 셈이다. 따라서 LSTM보다는 간단한 구조를 갖게 되었지만, Chung(2014) GRU의 성능은 LSTM에 뒤처지지 않으며 점차 다양한 분야에서 활용되어지고 있다고 하며, GRU구조는 아래와 같으며 이를 도식화한 것은 <그림 3>과 같다[17][18].

$$z_t = \sigma(W_z \bullet [h_{t-1}, x_t] + b_z) \quad (10)$$
$$r_t = \sigma(W_r \bullet [h_{t-1}, x_t] + b_r) \quad (11)$$
$$\widetilde{h_t} = \tanh(W_h \bullet [r_t \odot h_{t-1}, x_t] + b_h) \quad (12)$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t} \quad (13)$$

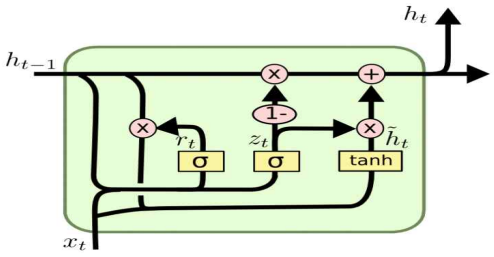

그림 3. Gated Recurrent Unit 도식화
Fig. 3. Gated Recurrent Unit structure

Ⅲ. 연구설계

3-1 데이터 수집

본 연구에서는 인공신경망을 적용한 분류 모델을 이용하여 [명사], [명사 + 형용사], [명사 + 형용사 + 동사], [모든 품사]의 4가지 품사 처리 방식과 RNN, LSTM, GRU 3가지 알고리즘을 적용한 총 12개의 모델의 성능을 비교하고자 한다. 악성 댓글은 주로 모욕적 발언, 선정적 발언, 정치적 발언 등 유형이 나타난다. 정치적 내용 댓글이 많을 경우 연구자의 주관에 상대적으로 많이 영향을 받는 것을 고려하여 정치적 내용이 상대적으로 적을 것이라 예상되는 연예 기사를 중점으로 댓글을 선별하고자 한다. 댓글의 수집은 '크롤링'을 이용

하였다. 선별한 10개의 기사는 [표 1]과 같으며, 분석을 위한 댓글이 최소 400개에서 최대 3,000개 이상까지 달린 기사를 위주로 선정하였다.

표 1. 댓글 수집을 위한 기사목록
Table 1. List of Articles for Collecting Comments

Title	Press	Comment	Ratio
I am the only person who is 'I live alone' ... Jeon Hyun-moo ♥ Han Hye-jin, 1 year 'Timeline'	Spotv news	784	5.87
Jeon Hyun-moo and Han Hye-jin "Break", “Good friend".	ytn	1747	13.07
Jeon Hyun-moo and Han Hye-jin "Break".. 'I live alone' short break	jtbc plus	2074	15.52
Jeon Hyun-moo and Han Hye-jin's breakup → "I live alone" break → Production team "Wait"	sports donga	2095	15.68
Jeon Hyun-moo and Han Hye-jin , Public love→One year break…'I live alone' Worrying	TV report	2900	21.70
'Break up' Jeon Hyun-moo · Han Hye-jin, falling into dilemma	nocut news	1124	8.41
'I live alone' Jeon Hyun-moo · Han Hye-jin, I left without a break	TV report	617	4.62
Jeon Hyeon-moo, Han Hye-jin, eventually break up "Good colleagues ..." I live alone "break"	Sports Seoul	1016	7.60
Jeon Hyun-moo and Han Hye-jin, One year break→'I live alone' break.."vacancy"	OSEN	600	4.49
Jeon Hyun-moo and Han Hye-jin, 'I live alone' Couple→ Break→Deny→ Three months as a colleague	TV report	405	3.03
Total	13,362		100

3-1 데이터 분류 및 처리

딥러닝을 이용한 악성 댓글 분류 모델 간의 성능 비교를 하기위해 악성댓글과 일반댓글을 분류해야 된다. 기존의 연구는 1명에 연구자의 주관성에 의존하여 악성댓글과 일반댓글을 분류해왔지만 본 연구에서는 연구자의 주관성 개입을 최소화하기위해 2단계에 거쳐 악성댓글과 일반댓글 분류를 실시하고자 한다. 첫 번째 단계로, 10개 네이버 뉴스 기사의 총 13,362개의 댓글을 한명의 연구자가 직접 읽어서 연구자의 주관에 따른 악성 댓글 556개를 1차 선별하였다. 두 번째 단계로 556개의 1차 선별된 악성 댓글을 3명의 연구자 중 2명 이상이 악성댓글로 동의할 경우 최종 악성 댓글로 조작적 정의하였으며 [표 2]와 같다. 본 연구에서는 데이터 품사 처리 과정에 있어 Python 3.7과 konlpy패키지의 형태소 분석기 중 하나인 okt를 사용하여 한글 형태소 분석을 실시하였다. 텍스트 데이터 품사 추출 방법은 총 4가지로 문장의 명사만 추출한 경우, 문장의 명사와 형용사만을 추출한 경우, 문장의 명사와 형용사, 동사를 추출한 경우, 조사를 포함한 모든 형태소를 추출한 경우로 분류하였다.

표 2. 수집된 댓글의 비율
Table 2. Ratio of Collected Comments

	General Comments		Malicious comments		Total
	Count	Ratio	Count	Ratio	Count
First	12,806	95.84	556	4.16	13,362
Second	13,098	98.02	264	1.98	13,362

Ⅳ. 인공신경망 성능 비교 분석결과

본 연구에서는 jupyter notebook과 keras패키지를 이용하였다. 신경망 은닉층의 뉴런은 32개로 고정하여 사용하였으며, 반복 학습 수는 500번을 수행하여 자료에 대해 충분히 학습시켰다. 학습된 모델은 성능평가는 [표 3]의 분류 성과 분석 테이블을 이용하여 가장 일반적으로 사용하는 Accuracy, Precise, Recall, f1-score 평가를 진행하였다. Accuracy는 정확도로 전체 분류 결과 중 True인 문항을 True로 선택한 정도와 False인 문항을 False의 비율의 합이다. Precise는 정밀도로 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율이다. Recall은 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율이다. f1-score는 Precision과 Recall의 조화평균으로 데이터 label이 불균형 구조일 때, 모델의 성능을 정확하게 평가할 수 있는 척도이다. 본 연구는 악성댓글과 일반댓글의 불균형을 2가지의 데이터셋을 구성하였다. 첫째, 13,362개의 전체 데이터셋과 둘째, 악성댓글과 일반댓글을 1대1 비율로 조정한 데이터셋이다. 2가지 데이터셋을 대상으로 4가지의 품사 처리 방법과 RNN, LSTM, GRU 3가지의 알고리즘을 조합한 모델 간의 성능 비교한 결과는 [표 4], [표 5]과 같다.

표 3. Classification 성과 분석 테이블
Table 3. Classification Performance Analysis Table

		Predicted Condition	
		True	False
True Condition	True	True Positive	False Positive
	False	False Negative	True Negative

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$f1 - score = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall}$$

표 4. 전체 데이터셋 : 분류 모델의 성능 비교
Table 4. All Data Set: Performance Comparison of Classification Model

	Accuracy				Precision				Recall				f1-score			
	RNN	LSTM	GRU	Avg.	RNN	LSTM	GRU	Avg.	RNN	LSTM	GRU	Avg.	RNN	LSTM	GRU	Avg.
Noun	80.27	90.82	90.31	87.13	2.15	3.77	3.70	3.21	19.57	28.26	19.57	22.47	3.87	6.65	6.23	5.58
Noun Adjective	82.18	88.40	91.52	87.37	2.94	4.14	4.20	3.76	36.17	25.53	23.4	28.37	5.44	7.12	7.12	6.56
Noun Adjective verb	88.21	93.08	94.87	92.05	5.88	11.49	9.09	8.82	22.00	20.00	32.00	24.67	9.28	14.60	14.16	12.68
All parts of speech	95.02	96.55	91.51	94.36	4.83	9.71	8.33	7.62	14.00	20.00	32.00	22.00	7.18	13.07	13.22	11.16
Average	86.42	92.21	92.05		3.95	7.28	6.33		22.94	23.45	26.74		6.44	10.36	10.18	

표 5. 1:1 데이터셋 : 분류 모델의 성능 비교
Table 5. 1:1 Dataset: Performance Comparison of Classification Model

	Accuracy				Precision				Recall				f1-score			
	RNN	LSTM	GRU	Avg.	RNN	LSTM	GRU	Avg.	RNN	LSTM	GRU	Avg.	RNN	LSTM	GRU	Avg.
Noun	57.14	76.19	75.24	69.52	56.36	74.55	74.07	68.33	59.62	78.85	76.92	71.80	57.94	76.64	75.47	70.02
Noun Adjective	60.38	80.19	74.53	71.70	58.93	76.27	82.05	72.42	63.46	86.54	61.54	70.51	61.11	81.08	70.33	70.84
Noun Adjective verb	68.87	77.36	77.36	74.53	67.27	71.88	78.00	72.38	71.15	88.46	75.00	78.20	69.16	79.31	76.47	74.98
All parts of speech	62.26	77.36	73.58	71.07	58.82	71.88	65.38	65.36	76.92	88.46	98.08	87.82	66.67	79.31	78.46	74.81
Average	62.16	77.78	75.18		60.35	73.65	74.88		67.79	85.58	77.89		63.72	79.09	75.18	

4-1 전체 데이터셋 성능 비교 결과

전체 데이터셋에서 각각의 모델들의 성능비교를 한 결과는 [표 6]과 같다. 정확도 순위의 경우 [모든 품사, LSTM], [모든 품사, RNN], [명사 + 형용사 + 동사, LSTM] 순으로 높게 나타났다. 품사처리 방법과 알고리즘의 정확도 평균의 경우 사용하는 품사가 많을수록 향상되는 것을 확인 할 수 있었으며, LSTM과 GRU이 RNN보다 상대적으로 높은 성능이 나타났다. 하지만 [모든 품사, RNN]의 조합이 95.02%의 정확도를 나타낸 것을 통해 전체 데이터셋에서 사용되는 품사에 따라 최적의 알고리즘의 차이가 있음을 확인하였다. f1-score까지 고려하였을 경우 12개의 악성 댓글 분류 모델 중 [모든 품사, LSTM]과 [명사 + 형용사 + 동사, GRU]가 가장 준수한 성능을 나타내는 것을 확인하였다.

표 6. 전체 데이터셋 : 분류 모델의 성능 순위
Table 6. All Data Set: Rank of Classification Model

	Ranking			
	Accuracy	Precision	Recall	f1-score
[Noun, RNN]	12	12	10	12
[Noun, LSTM]	7	9	4	9
[Noun, GRU]	8	10	10	10
[Noun + Adjective, RNN]	11	11	1	11
[Noun + Adjective, LSTM]	9	8	5	7
[Noun + Adjective, GRU]	5	7	6	7
[Noun + Adjective + Verb, RNN]	10	5	7	5
[Noun + Adjective + Verb, LSTM]	4	1	8	1
[Noun + Adjective + Verb, GRU]	3	3	2	2
[All parts of speech, RNN]	2	6	12	6
[All parts of speech, LSTM]	1	2	8	4
[All parts of speech, GRU]	6	4	2	3

4-2 1:1 데이터셋 성능 비교 결과

1:1 데이터셋에서 각각의 모델들의 성능 비교를 한 결과는 [표 7]과 같다. 정확도의 순위의 경우 [명사 + 형용사, LSTM], [명사 + 형용사 + 동사, LSTM], [모든 품사, LSTM] 순으로 나타났으며, LSTM을 적용한 모델이 모두 높은 순위가 나타난 것을 확인하였다. 품사처리 방법과 알고리즘의 정확도의 평균의 경우 [명사 + 형용사 + 동사], [명사 + 형용사], [모든 품사], [명사] 순으로 나타났으며, 전체 데이터셋과 평균의 순위는 차이가 있지만 명사만을 이용하는 경우가 가장 낮은 평균 순위를 나타낸 것은 동일하게 관찰되었다. 알고리즘별 정확도의 평균의 경우 LSTM, GRU, RNN 순으로 전체 데이터셋 과 동일한 순위를 나타내는 것을 확인하였다. 1:1 데이터셋에서 f1-score까지 고려하였을 때 f1-score 역시 정확도와 동일하게 [명사 + 형용사, LSTM], [명사 + 형용사 + 동사, LSTM], [모든 품사, LSTM] 순으로 높게 나타났고 3개의 모델이 가장 준수한 성능을 나타냈다.

표 7. 1:1 데이터셋 : 분류 모델의 성능 순위
Table 7. 1:1 Dataset: Rank of Classification Model

	Ranking			
	Accuracy	Precision	Recall	f1-score
[Noun, RNN]	12	12	12	12
[Noun, LSTM]	5	4	5	5
[Noun, GRU]	6	5	6	7
[Noun + Adjective, RNN]	11	10	10	11
[Noun + Adjective, LSTM]	1	3	4	1
[Noun + Adjective, GRU]	7	1	11	8
[Noun + Adjective + Verb, RNN]	9	8	9	9
[Noun + Adjective + Verb, LSTM]	2	6	2	2
[Noun + Adjective + Verb, GRU]	2	2	8	6
[All parts of speech, RNN]	10	11	6	10
[All parts of speech, LSTM]	2	6	2	2
[All parts of speech, GRU]	8	9	1	4

Ⅴ. 결론 및 한계점

본 연구에서는 인공신경망을 적용한 분류 모델을 이용하여 [명사], [명사 + 형용사], [명사 + 형용사 + 동사], [모든 품사]의 4가지 품사 처리 방식과 RNN, LSTM, GRU 3가지 분류알고리즘을 적용한 총 12개의 모델의 성능을 비교하였다. 인터넷의 발달로 인한 데이터의 급증과, 데이터를 저장할 수 있는 메모리 용량 증가, 처리기술의 발달에 따라 데이터를 분석하고 사회현상을 추적하는 빅데이터 시대에서 SNS에서 발생한 텍스트 데이터, 신문기사와 댓글, 다양한 인터넷 문서들은 기하급수적으로 증가할 것이며, 댓글 데이터는 문서의 수와 비례하게 기하급수적으로 증가 할 것이다. 정보통신망법 제44조의3항에 따르면 포털들은 모니터링 중 명예훼손성 게시글이 발생할 경우 자체적으로 제거할 수 있지만 깨끗한 인터넷 환경을 조성하고 악성 댓글 문제를 해결하기위해서는 AI를 이용한 악성 댓글을 분류하는 시스템을 개발이 필요 할 것이다.

본 연구의 시사점은 다음과 같다. 첫째, 다양한 품사 처리와 알고리즘을 이용하여 예측 모형들 간의 성능 비교를 통해 악성댓글 분류에 대한 최적 모형을 도출하고자 하였다. 본 연구에서는 [명사], [명사 + 형용사], [명사 + 형용사 + 동사], [모든 품사] 4가지의 품사 처리 방법과 RNN, LSTM, GRU의 3가지 알고리즘 이용하여 성능 비교를 수행하였으며, [모든 품사, LSTM], [모든 품사, RNN], [명사 + 형용사 + 동사, GRU]조합의 순으로 정확도가 높게 나타났다. 한글 자연어 처리 분야는 지속적으로 연구가 진행되는 분야이며, 다양한 알고리즘과 품사 처리 조합을 이용하여 최적의 모형을 이용하는 시도는 다른 연구에서도 참고자료가 될 수 있을 것으로 사료된다. 둘째, AI를 이용한 한글 악성텍스트 분류와 차단을 위한 기초자료로서 시도한 점이다. 구글은 게시글의 악성지수를 계산해주는 Perspective를 개발하고 있으며, 인스타그램은 AI를 이용한 악성 댓글차단기능을 개발하고 있다. 한글 자연어처리 분야의 경우 '파파고', '플리토' 등 번역 분야에서 AI도입을 시도하고 있지만, 텍스트의 악성여부와 같은 상대적으로 주관성이 영향을 강하게 미칠 수 있는 분야에 대한 연구는 타 분야에 비해 실무적으로도 부족하다고 볼 수 있다. 본 연구는 한글 악성텍스트 분류와 차단을 위한 기초자료로서 그 의미가 있다고 사료된다. 본 연구에서 악성 댓글 분류 모델간의 성능 비교를 수행하는 과정에 있어서 최대한 객관적으로 분석하고자 하였으나, 현실적인 한계로 인해 다음과 같은 한계점을 내포하고 있다. 첫째, 상대적으로 부족한 댓글 데이터이다. 본 연구에서는 "전현무, 한혜진 결별"관련된 네이버 뉴스 10개 기사의 13,362개의 댓글을 활용하였지만, 해당주제에 관련된 모든 댓글데이터를 활용하지 못한 한계점이 존재한다. 두 번째로 실제 데이터셋에서의 다소 높지 않은 분류 모델의 성능이다. 13,362개의 댓글 중 악성댓글은 264개로 1.98% 비중을 차지하고 있었다. 하지만 분류모델의 성능이 가장 높은 조합인 [모든 품사, LSTM]의 정확도는 96.55%로서 성능이 다소 높지 않은 것으로 보여졌다. 향후 연구에서는 분류모델의 성능 향상을 위해 워드 임베딩을 도입하여 분류모델의 성능 향상을 시도할 필요가 있다.

감사의 글

본 논문은 김진우의 2019년도 석사 학위논문에서 데이터를 활용하여 재구성 및 발췌 정리하였음

참 고 문 헌

[1] Korea National Police Agency,[Internet], Available: https://www.police.go.kr/portal/main/contents.do?menuNo=200550

[2] Status of Internet comment regulation and legislative review tasks, National Assembly Legislative Research Center, 2018. 04. 30.

[3] [Internet], Available: https://opendict.korean.go.kr/dictionary/view?sense_no=739029&viewType=confirm

[4] M. J. Kim, "Regulating Hate Speech at ilbe.com? A Conceptual Analysis of Online Hate Speech", *Journal of Media Law, Ethics and Policy Research,* Vol 13, No. 2, pp. 131-163, December, 2014.

[5] S. S. Hong, "Articles : Regulating Hate Speech: A Legal Strategy to Promote Freedom of Speech and to Protect the Right of Minorities", *Law and Society,* Vol 50, No. 0, pp. 287-336, 2015.

[6] H. Y. Park, "The Constitutional Study on the Hate Speech", *Korean Comparative Public Law Association,* Vol 16, No. 3, pp. 137-169, August, 2014.

[7] Y. I. Kim, "The Characteristics of Malicious Comments : Comparisons of the Internet News Comments in Korean and English", *JOURNAL OF THE KOREA CONTENTS ASSOCIATION,* Vol. 19, No. 1, pp. 548-558, January , 2019.

[8] H. S. Yang, "Hate speech toward specific regions in Korea : Content analysis of comments posted to crime news stories in Naver", *Korean Journal of Journalism & Communication Studies,* Vol. 62, No. 6, pp. 7-36, December, 2018.

[9] C. Y. Jeon, "An Exploratory Study on the Hate Speech Restriction Decisions by the Korea Communications Standards Commission - Actual Status and Regulation of Online Hate Speech", *Journal of Broadcasting and Telecommunications Research* , pp. 70-102, October, 2018.

[10] J. H. Hong, "Online Hate Speech Diffusion Network Analysis : Issue-Specific Diffusion Patterns, Types and Intensity of Verbal Expression on Online Hatred", *Korean Society For Journalism And Communication Studies*, Vol. 60, No. 5, pp. 145-175, October, 2016.

[11] Lee, H. S., Lee, H. R., Park, J. U., & Han, Y. S. "An abusive text detection system based on enhanced abusive and non-abusive word lists." *Decision Support Systems*, Vol. 113, pp. 22-31, 2018.

[12] Aiyar, S., & Shetty, N. P.N-gram assisted Youtube spam comment detection. *Procedia computer science*, Vol. 132, pp. 174-182, 2018.

[13] M. S. Kim, "A Design and Implementation of Malicious Web Log Identification System by Using SVM", *Korean Institute of Information Scientists and Engineers*, pp. 285-289, October, 2006.

[14] J. J. Hong, "A Malicious Comments Detection Technique on the Internet using Sentiment Analysis and SVM", *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 20, No. 2, pp. 260-267, February, 2016.

[15] C. Y. Park, ”Data mining using R” , pp. 95-107, 2011.

[16] Mikolov, Tomas, et al. "Recurrent neural network based language model." Interspeech. Vol. 2, 2010.

[17] [Internet], Available: http://dprogrammer.org/rnn-lstm-gru

[18] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555. 2014.

김진우(Jin Woo Kim)

2017년 : 고려대학교 정보통계학과 (학사)

2017년~현 재: 연세대학교 정보대학원 ICT·콘텐츠트랙 (석사과정)
※관심분야 : 빅데이터 분석, 소셜 데이터, 미디어 정책 등

조혜인(Hye In Jo)

2018년 : 순천향대학교 컴퓨터공학과 (학사)

2018년~현 재: 연세대학교 정보대학원 ICT·콘텐츠트랙 (석사과정)
※관심분야 : 텍스트 마이닝, 데이터 아키텍처, 데이터 분석 등

이봉규(Bong Gyou Lee)

1988년 : 연세대학교 상경대학 경제학과 (학사)
1992년 : Cornell University, Dept. of CRP (MS)
1994년 : Cornell University, Dept. of CRP (Ph.D)

1997년~2005년: 한성대학교 공과대학 정보전산학부 교수
2016년~2017년: 연세대학교 정보대학원 원장
2005년~현 재: 연세대학교 정보대학원 ICT·콘텐츠트랙 교수
2018년~현 재: 연세대학교 학술정보원 원장(CIO, CPO)
※관심분야 : 디지털 트랜스포메이션 기술 및 전략 등