韓國IT서비스學會誌
第15卷 第3號
2016年 9月, pp.147-155

Journal of Information Technology Services
http://dx.doi.org/10.9716/KITS.2016.15.3.147

딥러닝을 활용한 웹 텍스트 저자의 남녀 구분 및 연령 판별 : SNS 사용자를 중심으로

박찬엽* · 장인호** · 이준기***

Authorship Attribution of Web Texts with Korean Language Applying Deep Learning Method

Chan Yub Park* · In Ho Jang** · Zoon Ky Lee***

■ Abstract ■

According to rapid development of technology, web text is growing explosively and attracting many fields as substitution for survey. The user of Facebook is reaching up to 113 million people per month, Twitter is used in various institution or company as a behavioral analysis tool. However, many research has focused on meaning of the text itself. And there is a lack of study for text's creation subject. Therefore, this research consists of sex/age text classification with by using 20,187 Facebook users' posts that reveal the sex and age of the writer. This research utilized Convolution Neural Networks, a type of deep learning algorithms which came into the spotlight as a recent image classifier in web text analyzing. The following result assured with 92% of accuracy for possibility as a text classifier. Also, this research was minimizing the Korean morpheme analysis and it was conducted using a Korean web text to Authorship Attribution. Based on these feature, this study can develop users' multiple capacity such as web text management information resource for worker, non-grammatical analyzing system for researchers. Thus, this study proposes a new method for web text analysis.

Keyword : Deep Learning, Convolutional Neural Networks, Authorship Attribution, Web Text Analysis, Facebook

Submitted : May 11, 2016 1st Revision : July 28, 2016 Accepted : August 2, 2016

* 공개SW역량프라자 연구원
** 연세대학교 정보대학원 석사과정
*** 연세대학교 정보대학원 교수, 교신저자

1. 서 론

최근 웹 2.0 기술의 발달과 스마트 기기의 확산으로 소셜 네트워킹 서비스(Social Networking Service 이하 SNS)가 빠르게 사용자층을 넓히고 있다. SNS는 자신의 취향과 활동을 공유하거나, 타인의 취향과 활동을 관찰하고자 하는 사람들의 공동체를 위한 온라인 사회관계 형성을 주도하고 있다(IWGDPT, 2008). 이러한 SNS의 발전과 더불어 국내 SNS 이용자 수도 기하급수적으로 증가하고 있다.

실제로 2015년 9월 기준으로 국내 페이스북 월간 이용자수는 1,130만 명에 이르고 있으며 이를 바탕으로 여러 기업과 기관에서 SNS 분석을 여론 및 소비자 분석 도구로 사용하고 있다. 이 같은 SNS 사용 목적의 다양화와 이용자의 증가는 SNS의 학문적 활용 방안에 대한 관심으로 이어졌다. Kang and Lee(2014)에 따르면 트위터의 경우 2009년부터 2014년 4월까지 국내 트위터 관련 논문이 총 539편이 존재하며 관련연구가 53개 학문분야에서 다양하게 이루어져 있다고 밝혔다. 이 연구를 통해 트위터 관련 연구가 선거나 정치 관련 주제로 가장 많이 다루어졌다는 것을 알 수 있다.

이처럼 SNS에 대한 연구는 국내에서도 활발하게 이뤄지고 있지만, 아직 그 활용도에 대한 것들이 주를 이루고 있다. 더불어 SNS 상의 텍스트를 작성한 유저를 식별하여 경영 및 마케팅 그리고 인문·사회적으로 유의미함을 찾는 연구는 아직 미미한 실정이다. Bhargava(2013)에 따르면 해외의 경우 익명의 SNS 텍스트를 저자판별하고 이를 다양한 목적으로 활용하기 위한 연구가 진행되고 있다.

Bhargava(2013)는 유사도 기반 방법을 활용하여 '트윗' 작성자를 판별하는 연구를 통해 트위터 상의 짧은 텍스트를 과학적 데이터 분석의 대상으로 활용했다. Mikolov(2013) 또한 트위터 텍스트를 이용하여 저자판별을 시도했는데, 여기서는 그리스어 '트윗' 12,973개를 데이터 세트로 하고 글자 N-gram과 단어 N-gram 정보를 결합한 AMNP (Author's Multilevel N-gram Profile) 13이라는 자질을 이용해 멀티클래스 서포트 벡터(Multiclass SVMs) 기법으로 저자 판별을 시도했다.

이처럼 해외에서는 SNS 상의 짧은 텍스트 데이터의 기본을 파악하고 이를 과학적 데이터 분석으로 사용하는 연구가 활발하게 진행되고 있다.

하지만 국내 저자판별 연구의 경우 한국어 텍스트만으로 저자를 파악하는 연구는 총 5건으로 이 중 웹 텍스트를 대상으로 한 저자 판별 연구는 2건에 불과하다. Han(2009)의 경우 한 신문사의 기사를 데이터로 활용하여 통계적, 정량적 방법을 한글 저자 판별에 적용하는 연구를 진행했으며, Choi(2015)은 블로그의 영화 리뷰를 중심으로 기계학습을 통한 저자 판별 연구를 시도 했는데, 두 연구 모두 한글에서도 성공적으로 저자 판별이 가능하다는 의의를 남겼으나 한글의 문법적 특성에 따른 분석의 어려움과 대상 데이터 확보의 문제를 나타내고 있다.

따라서 본 연구는 딥러닝 기법을 활용하여 한글의 문법적 특성에 따른 분석의 어려움을 최소화하고 웹 상의 텍스트를 대량으로 수집하여 분석에 활용하였고, 그 결과를 검증한다.

본 연구의 구성은 다음과 같다. 제 2장에서 저자판별과 딥러닝에 대해서 알아보고, 제 3장에서는 연구의 주된 기법인 컨볼루셔널 신경망(Convolutional Neural Network)에 대해 소개한다. 그리고 제 4장에서 제안한 연구 방법을 기술하며, 제 5장에서는 연구 방법에 따른 결과를 나타낸다. 마지막으로 제 6장에서는 이 연구의 한계점과 향후 연구방안에 대해 기술할 것이다.

2. 연구 배경

저자판별은 작자가 무기명으로 되어있거나 작자의 진위가 논쟁이 되고 있는 저작물에 작자를 할

당하는 작업이다(Han, 2009).

저자판별 연구는 유사도 기반 방법과 기계학습 방법으로 나눌 수 있다(Stamatatos, 2009). 유사도 기반을 통한 저자 판별 연구는 기존의 저자가 알려진 문서들과 무기명 문서의 특징을 측정하여 이를 여러 거리 계산 방법으로 가장 거리가 가까운 문서를 찾아 저자를 추정하는 방법이다(Abbasi and Chen, 2008; Argamon et al., 2009). 그리고 기계학습 방법은 저자가 알려진 문서와 무기명 문서의 특징을 측정하고 이를 바탕으로 머신 러닝 분류기를 구성하여 무기명 문서를 분류해 내는 방법이다(Zheng et al., 2006; Abbasi and Chen, 2008).

유사도 기반 방법은 효과적인 특징 추출 및 추상화 방법과 거리 계산방식에 대해 집중하지만, 기계학습 방법은 특징 추출과 분류기 선정 및 분류기 파라미터 최적화에 집중한다.

최근 기계학습 분야에서는 딥러닝 기법의 발달로 해당 분야의 지식 없이 데이터로부터 자동으로 특징을 추출해내는 연구가 시작되고 있다. 특히 특징 추출기와 분류기를 대규모 신경망으로 통합하여 학습함으로써 기존의 기계학습에 비해 비약적인 성능 향상을 이루었다.

딥러닝 기법 중 하나인 컨볼루셔널 신경망(Convolutional Neural Network, 이하 CNN)과 리커런트 신경망(Recurrent Neural Network)의 경우 최근 영상 인식에 널리 활용되고 있으며 리커런트 신경망의 일종인 LSTM(Long Short Term Memory)는 필기체 인식이나 음성인식에 성공적으로 적용되고 있는 등 여러 가지 딥러닝 기법들이 계속해서 발전하며 다양한 분야에서 적용되고 있다.

본 연구에서는 기존의 국내 저자판별 연구에서 활용한 유사도 기반 방법이 아닌, 기계학습 기법 중 새롭게 각광 받고 있는 CNN을 한글 텍스트 저자판별에 적용한다. 또한 데이터 확보에 어려움을 겪었던 기존 연구를 보완하여 현실의 웹 텍스트를 확보하고 이를 기반으로 한글의 자연어처리를 최소화한 한국어 텍스트 저자판별 방법론을 제시하는 것을 목적으로 한다.

이 연구는 딥러닝 기법을 한글 텍스트 대상으로 하여 저자판별의 연구에 적용한 것으로, 향후 딥러닝 연구가 다양한 분야에 접목이 가능하게끔 해주는 방향성을 제시할 수 있을 것이다.

3. 이론적 배경

3.1 인공신경망

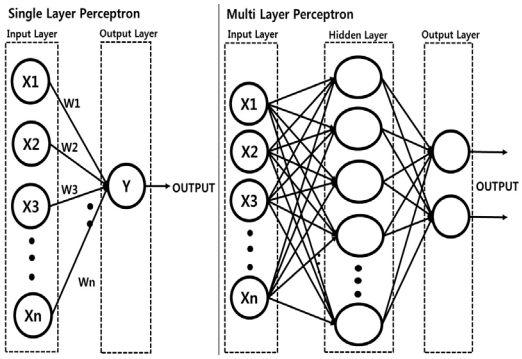
인공신경망은 기계학습의 한 종류로, 사람의 뇌가 학습하는 방법을 모사하여 컴퓨터로 모델링하는 소프트웨어적 방법이다.

사람의 뇌신경에는 뉴런이라는 신경 세포가 망으로 구성되어 있다. 뉴런의 기능은 자극을 받았을 경우 전기를 발생시켜 다른 세포에 정보를 전달하는 것으로, 자극은 뉴런의 수상 돌기에 전달되고 축색 돌기를 통해 말단으로 전달된다. 이 때, 서로 다른 수상돌기와 축색 돌기의 말단이 만나는 곳인 시냅스에서 신경전달 물질이 생성된다.

이런 뉴런의 작용을 모사한 것이 인공 신경망으로, 인공 신경망에서는 뉴런과 같은 기본단위를 퍼셉트론이라고 하는데 입력층과 출력층의 두 개의 노드로 이루어진 가장 단순한 형태를 단층 퍼셉트론이라고 한다.

단층 퍼셉트론은 입력노드로 들어오는 값에 가중치를 고려하여 출력노드에 전달할 값을 결정한다. 하지만 단층 퍼셉트론은 이진 분류기로서 그 한계에 부딪히는데, 이를 개선한 것이 단층 퍼셉트론을 여러 개 연결한 다층 퍼셉트론이다.

다층 퍼셉트론은 여러 개의 단층 퍼셉트론을 합친 모델로 입력층과 은닉층, 출력층으로 구성되어 있다. 다층 퍼셉트론은 입, 출력층의 중간에 은닉층을 삽입하여 선형 분리가 가능하도록 하며, 입출력 특성을 비선형화 함으로써 네트워크 능력을 향상시켜 단층 퍼셉트론의 여러 단점을 보완할 수 있다.



〈Figure 1〉 Single-Layer Perceptron and Multi-Layer Perceptron Structure

3.2 컨볼루셔널 신경망

CNN은 컨볼루션(Convolution) 필터로 계산하여 특징을 추출하는 층과 서브샘플링(Subsampling)으로 계산량을 효율적으로 축소시키는 풀링(Pooling) 층으로 구성되어 있다. 컨볼루션을 수행하는 것은 입력층의 홀수 배 크기인 커널(Kernel)의 중심이 입력 프레임의 픽셀에 놓인 상태에서 입력 프레임과 커널이 겹쳐진 부분들만 계산해 출력값을 만드는 과정이다. 커널과 입력 프레임을 겹쳐 컨볼루션을 진행하는 방법은 3가지가 존재하는데, 본 연구에서는 커널을 입력 프레임에 완전히 겹치게 하여 출력 프레임을 축소시키는 것을 기본으로 했다.

픽셀이 매우 높은 경우 컨볼루션을 수행하는데 많은 시간이 소요 되는데, 서브샘플링 단계는 이 계산량을 효과적으로 줄이는 방법으로, 4개 픽셀의 평균값을 결과 프레임으로 출력하는 평균 풀링(Mean Pooling)과 4개 픽셀 중 가장 높은 값을 결과 프레임으로 출력하는 맥스 풀링(Max Pooling) 방법이 대중적으로 사용되고 있다. 본 연구에서는 이 중 성능이 우수한 맥스 풀링 방법을 사용했다.

3.3 텍스트 분류기로서 CNN

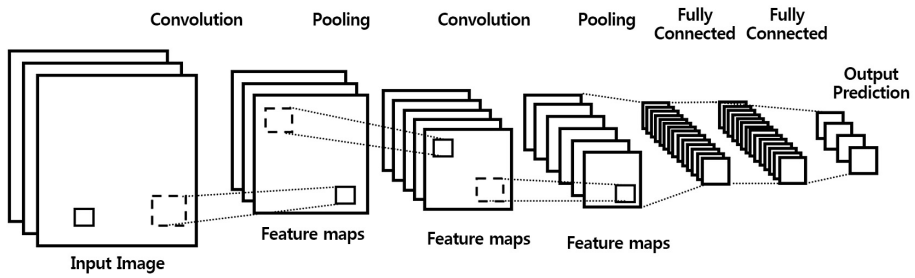
Zhang et al.(2015)은 CNN을 활용하여 문자 수준(Charater-level) 텍스트 데이터를 분류하는 연구를 진행했다. 한글 자모에 해당하는 영어 알파벳과 공백, 일정 특수문자를 분류하고 이를 70개의 세로 프레임, 1,024개의 가로 프레임의 벡터 이미지로 전환하는 방법을 제안했다.

이 같은 텍스트의 통계적 문자 수준 분석은 데이터의 수가 많을수록 정확도가 높아지며, 특히 Zhang et al.(2015)이 제안한 방법은 교착어 데이터 분석이 복잡한 문법 처리에서 자유로울 수 있다는 것을 증명하였는데, 이 연구를 통해 교착어인 한글에도 문자 수준 분석을 적용하는 것이 매우 긍정적이라는 것을 확인하였다.

4. 연구 방법

4.1 데이터 입력

본 연구에서는 텍스트의 연속된 값을 입력층으로 사용한다. 입력층의 차원을 최소화하기 위해 24자의 기본자모 수준 데이터와 띄어쓰기, 특수 문자 일부를 포함해 총 50개의 글자로 입력층 데이



〈Figure 2〉 Convolutional Neural Networks Structure

터를 제한했으며 포함하는 글자는 <Figure 3>과 같다.

〈Figure 3〉 Including Letters and Special Characters

모든 복합자모는 기본 자모의 합성으로 구성하였는데, 예로 ㄲ(쌍기역)은 ㄱㄱ, 애는 ㅏㅣ로 표기하였다. 이 방법으로 자모의 구성을 줄여 입력층의 차원을 최소화 할 수 있었다.

각 자모 및 특수문자의 연속된 값은 최대 1,024자로 제한하였다. 1,024자는 본 연구에서 사용한 오픈 소스인 크레페(Crepe)의 기본 설정이며, 이는 일정 텍스트의 양을 확보함으로써 짧은 글에서 생기는 정보 부족을 최소화 할 수 있는 단위이다.

위 기준을 바탕으로 생성되는 본 연구의 입력층은 가로 50, 길이 1,024인 프레임의 기본 단위이다.

4.2 연구 모형 및 환경

본 연구에서는 텍스트를 이해할 CNN을 6층의 컨볼루션 층과 3층의 풀리 커넥티드 층(Fully-Connected Layer), 총 9개 층으로 구성했다. 입력층은 앞서 구성한 50×1,024 크기의 프레임을 가지며 각 층 역시 1,024 프레임으로 구성하였다. 풀리 커넥티드 층의 마지막 층은 출력층으로 문제마다 적절한 클래스 개수를 조절하여 결과를 확인할 수 있다.

분석 툴은 페이스북이 공개한 딥러닝 플랫폼인 토치7(Torch7)을 활용한 오픈 소스 크레페를 사용하였으며, 분석 머신은 CPU : Xeon E5, GPU : Ndivia GTX 750i(cuda processor 680 uits), RAM : 16G DDR, OS : Ubunto 14.04 LTS로 구성하였다.

반복 학습은 5,000회, 20번 반복하였으며 드랍아웃(Dropout)은 풀리 커넥티드 층에서 작동하는 크레페의 기본 설정을 따랐다.

4.3 데이터 수집

본 연구에 사용된 데이터는 페이스북 내에 성별과 연령을 프로필에 공개한 사용자의 포스트를 수집하여 전처리 하였다. 수집도구는 R 3.2.2 on Window7 이며, 관련 패키지는 Rwebdriver와 Rvest를 사용하였다.

Rwebdriver는 클라이언트 핸들링 인터페이스인 셀레니움(Selenium)을 R에서 사용할 수 있게 작성된 패키지이고, Rvest는 DOM 문서의 전처리를 유용하게 해주는 패키지이다.

사용자 아이디를 확보하기 위해 국내 페이스북 페이지 중 '좋아요' 수 상위 20개 페이지의 최근 10개 포스트에 댓글 및 '좋아요'를 수행한 아이디를 페이스북 API를 이용해 수집하였다. 이 후, Rwebdriver와 Rvest를 활용하여 페이스북 이용자의 프로필 페이지를 파싱(Parsing)하여 연령과 성별이 공개되어 있는지를 확인하고 공개가 확인된 아이디만 따로 저장 하였다.

이렇게 저장된 아이디의 2015년 포스트만을 분석 대상으로 삼았으며, 이보다 과거의 데이터는 연령대 구분에 혼선이 있을 수 있기 때문에 제외하였다.

5. 연구 결과

5.1 대상 데이터

<Table 1>은 페이스북에서 수집한 데이터 중 생년월일과 성별 기록을 프로필에 공개한 총 20,187명의 아이디를 성별, 연령별로 나타낸 것이며, 이들의 2015년 1년간의 포스트를 데이터로 사용하였다.

〈Table 1〉 a Number of Collected ID by Age and Sex

	10s	20s	30s	Over 40s
Male	679	4,347	4,597	4,326
Female	532	3,873	1,185	648

60대의 경우 남성과 여성의 불균형이 두드러졌는데, 60대 남성은 공개적 활동을 통해 장문의 포스트를 많이 작성하는 경향을 보였으나, 60대 여성은 대상 아이디와 포스트 작성 수가 적었다. 따라서 60대 이상의 경우 데이터의 균질성을 확보할 수 없어 대상 데이터에서 제외했다.

10대 데이터 또한 샘플 수가 부족하며 확보된 샘플의 포스트수도 적어 대상 데이터로 사용하지 않았다. 이를 통해 CNN의 출력층은 최종적으로 20대 남·여, 30대 남·여, 40대 이상 남·여의 총 6개 집단으로 구성되었다.

수집한 아이디의 페이스북 타임 라인에서 총 527,172개의 포스트를 확보하였으며 전처리를 통해 3,000자 이상, 100자 미만의 포스트는 제거하였다. 또한 기본 자모로 변환한 길이가 1,024를 초과한 경우 글자를 자르고 새로운 포스트로 가정하여 샘플을 확보하였다. 이러한 방법으로 확보한 포스트 샘플은 <Table 2>와 같다.

〈Table 2〉 the Number of Posts by Age and Sex

	20s	30s	Over 40s
Male	146,286	135,869	99,802
Female	192,832	58,086	50,320

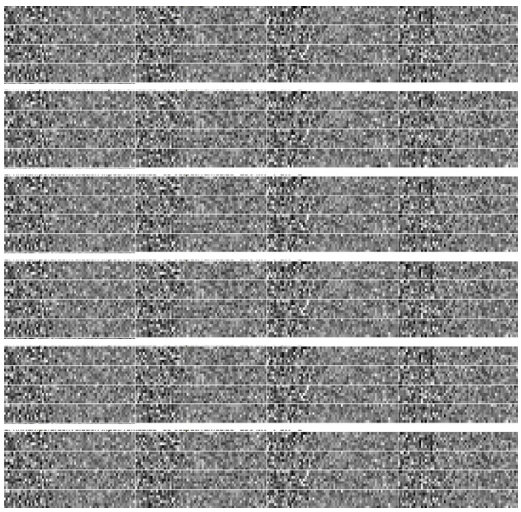
40대 이상 여성의 포스트 샘플 수가 약 5만 개로 6개의 집단 중 그 수가 가장 적어 균질한 분석을 위해 분석에 활용할 각 집단의 포스트 샘플 수를 5만 개로 설정했으며, 나머지 5개의 집단(20대 남·여, 30대 남·여, 40대 이상 남)은 5만 개의 포스트 샘플을 무작위 추출하여 최종 데이터 셋을 30만 개 구축하였다.

6개 집단 모두 학습 셋은 45,000개, 테스트 셋은 5,000개로 구성하여 분석을 진행했다.

5.2 분류 결과

분석에 활용한 오픈 소스 크레페는 학습과정 중 각 층의 가중치 상황을 이미지로 표현해 준다.

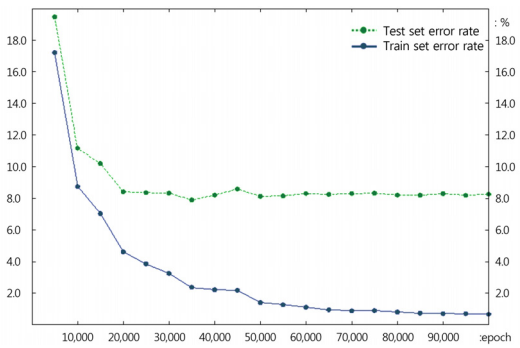
<Figure 4>는 반복 횟수에 따른 첫 번째 층의 이미지로 위에서부터 각각 10,000회, 50,000회, 100,000회, 150,000회, 200,000회, 225,000회 반복 학습 후의 이미지를 나열한 것이다.



〈Figure 4〉 Iterative Learning of the First-Layer

<Figure 4>의 픽셀 각각은 입력 프레임과 커널 간의 계산된 가중치를 의미하고 색이 회색에 가까우면 가중치가 0에 가까운 것을 의미하는데, 50,000회 이후로는 가중치 변화가 거의 없었다.

<Figure 5>에서 점선은 테스트 셋의 에러율을, 실선은 학습 셋의 에러율을 표시한 것이다.



〈Figure 5〉 Error Rate of Test Set and Train Set

50,000회를 기점으로 테스트 셋의 에러율이 약

8% 내외를 유지하고 있다. 이는 빈도 정보 기법을 이용하여 160개의 칼럼을 대상 데이터로 분석한 Han(2009)의 연구에서 보인 93% 정확도와 큰 차이가 없었다.

6. 결 론

6.1 시사점

저자 판별 연구가 다양한 방법으로 발전을 거듭하였지만 한글 웹 텍스트를 대상으로 한 저자 판별 연구는 미흡한 실정이었다.

본 연구는 한국 페이스북 페이지 '좋아요' 수 상위 20개의 페이지 내 최근 10개의 포스트에 '좋아요'와 댓글을 단 아이디 중 성별, 연령 정보가 입력된 사용자의 2015년 포스트를 분석용 데이터로 사용했다.

그리고 한글 웹 텍스트의 특징을 충실히 반영하여 저자 판별을 수행했다는 점, 기계학습에서의 새로운 패러다임인 딥러닝 알고리즘을 저자 판별 분야에 텍스트 분류기로 적용하고 검증하였다는 점에서 그 의미를 찾을 수 있다.

본 연구는 연구적, 실무적으로 시사하는 바가 있는데, 연구적 시사점으로는 첫 번째로 한글 저자 판별에 딥러닝 기법 중 하나인 CNN을 사용함으로써 한글의 문법적 특성에 따른 데이터 전처리의 복잡성과 분석의 어려움을 최소화하고자 했다는 것이다. 그리고 두 번째로는 본 연구가 기본 자모 단위의 통계적 분석으로도 충분히 효과적인 결과를 도출할 수 있다는 것을 확인함으로써 한글의 복잡한 문법체계로 제한적인 연구를 진행할 수 밖에 없었던 국내 웹 텍스트 연구에 새로운 대안을 제시하였다는 것이다.

실무적 시사점으로는 경영 관점에서 활용하기 좋은 페이스북의 데이터를 대상으로 하여 연령과 성별을 분류 기준으로 삼았고 기존의 텍스트 전체에 대한 분석 대신 연령대별 및 성별 분류를 통해 각 텍스트에 대한 분석을 진행할 수 있는 초기 연구를 진행했다는 것인데, 이는 웹텍스트를 경영 정보 자원으로써 활용했다는 점에서 실무적 의의를 갖고 있다.

본 연구는 저자 판별 연구에 있어 딥러닝 알고리즘인 CNN을 활용하여 검증하였고 텍스트 분석 방법으로의 활용 가능성에 대해 확인함으로써 향후 다양한 분야로의 적용을 확대해 가는 단초가 될 것이다.

6.2 한계점 및 향후 연구 방향

본 연구는 한글 웹 텍스트 데이터 분석에 있어 교착어의 문법적 특성을 극복하고 이전의 행동기반 고객군 추정으로 진행된 연구에서 벗어나 경영정보 자원으로써의 활용 안을 제시했다. 하지만 페이스북 이외의 다른 SNS 상의 웹 텍스트에 검증해보지 못했다는 점과 데이터가 부족한 다른 연령대를 포함하지 못했다는 점, 그리고 딥러닝 알고리즘의 블랙박스 현상으로 인해 세대와 연령의 분류에 대해 이해할 수 있는 변수를 찾기 힘들다는 한계점을 가지고 있다.

따라서 차후 연구에서는 현재 성별과 연령대가 표시된 웹 텍스트를 확보 할 수 있는 다른 데이터 소스가 거의 없다는 점에서 지속적인 말뭉치 구축이 필요해 보인다. 또한 미래의 고객으로서 트렌드에 민감한 10대의 웹 텍스트 사용을 추적하는 연구를 병행한다면 경영정보로서 더욱 가치 있는 연구가 될 것이다.

References

Abbasi, A. and H. Chen, "Writerprints : A Stylometric Approach to Identity-level Identification and Similarity Detection", *ACM Transactions on Information Systems*, Vol.26, No.2, 2008.

Argamon, S., M. Koppel, J.W. Pennebaker, and J. Schler, "Automatically Profiling the Au-

thor of an Anonymous Text", *Communications of the ACM*, Vol.52, No.2, 2009, 119-123.

Bhargava, M., P. Mehndiratta, and K. Asawa, "Stylometric Analysis for Authorship Attribution on Twitter", *BDA*, Vol.8302, 2013, 37-47.

Choi, J.M., "Authorship Attribution of Korean Texts Using Machine Learning Methods : A Study on Movie Reviews on Blogs", Yonsei University Master's thesis located, 2015.

(최지명, "기계학습을 활용한 한국어 텍스트 저자판별", 연세대학교 석사학위논문, 2015.)

Han, N.R., "Authorship Attribution in Korean Using Frequency Profiles", *KJCS*, Vol.20, No.2, 2009, 225-241.

(한나래, "빈도정보를 이용한 한국어 저자판별", *인지과학회지,* 제20권, 제2호, 2009, 225-241.)

IWGDPT, "Report and Guidance on Privacy in Social Network Services : Rome Memorandum", 2008. Available at http://www.datenschutz-berlin.de/attachments/461/WP_social_network_services.pdf(Downloaded June 15. 2015).

Kang, B.I. and J.Y. Lee, "A Bibliometric Analysis on Twitter Research", *Journal of the Korean Society for Information Management,* Vol.31, No.3, 2014, 293-311.

(강범일, 이재윤, "트위터 관련 연구에 대한 계량정보학적 분석", *정보관리학회지*, 제31권, 제3호, 2014, 293-311.)

Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", 2013. Available at https://arxiv.org/pdf/1301.3781.pdf(Downloaded June 12. 2015.)

Park, C.Y, "Korean Authorship Attribution from Web Texts Using Machine Learning Methods-Facebook post", *Yonsei University Master's thesis located*, 2015.

(박찬엽, "기계학습을 활용한 한국어 웹 텍스트 저자판별(성별, 연령별) : 페이스북 사용자를 중심으로", 연세대학교 석사학위논문, 2015.)

Stamatatos, E., "A Survey of Modern Authorship Attribution Methods", *Journal of the American Society for Information Science and Technology*, Vol.60, No.3, 2009, 538-556.

Zhang, X., J. Zhao, and Y. LeCun, "Character-Level Convolutional Networks for Text Classification", *Advances in Neural Information Processing Systems*, Vol.28, 2015.

Zheng, R., J.X. Li, H.C. Chen, and Z. Huang, "A Framework for Authorship Identification of Online Messages : Writing-style Features and Classification Techniques", *Journal of the American Society for Information Science and Technology*, Vol.57, No.3, 2006, 378-393.

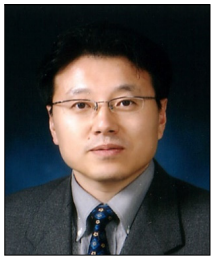
About the Authors

Chan Yub Park (mrchypark@gmail.com)

Chan Yub Park is currently a senior researcher at Open Source Software Competency Plaza. He received his master degree in Management Information Systems from graduate school of information, Yonsei university in 2016. His current research interests include text mining, machine translation, and etc.

In Ho Jang (norex@yonsei.ac.kr)

In Ho Jang is currently a master's course in MIS from graduate school information, Yonsei University and B.S. in Mechanical Engineering from Chonnam National University in 2016. His current research interests include data analysis, IoT, big data analysis, and etc.



Zoon Ky Lee (zoonky@gmail.com)

Professor Zoon Ky Lee is currently a Professor at Graduate School of information, Yonsei University. He received his B.S. degree in Computer Science and Statistics from Seoul National University and a Ph.D. degree in Management Information Systems from University of Southern California in 1994. His current research interests include Open Innovation, Service Science Big Data Applications, and E-Transformation