Information Systems Review
Vol. 21, No. 1 February 2019 http://dx.doi.org/10.14329/isr.2019.21.1.023

머신 러닝을 활용한 회사 SNS 메시지에 내포된 심리적 거리 추출 연구

A Study on the Extraction of Psychological Distance Embedded in Company's SNS Messages Using Machine Learning

이 성 원 (Seongwon Lee) 아주대학교 의과대학 의료정보학과 연구강사
김 진 혁 (Jin Hyuk Kim) 단국대학교 공과대학 컴퓨터과학과 학부생, 교신저자

요 약

소셜 네트워크 서비스(이하 SNS)는 회사의 마케팅 채널로 적극 활용되고 있으며, 회사들의 고객층에 적합한 내용과 어조를 활용하여 주기적으로 SNS 메시지를 작성하는 등 활발한 마케팅을 펼치고 있다. 본 논문에서는 이제까지 간과되었던 SNS 메시지에 내포된 심리적 거리에 초점을 맞춰 전통적인 코더를 활용한 내용 분석(content analysis)과 자연어 처리 기법 및 머신 러닝 방법을 혼합하여 심리적 거리를 측정하는 분석 방법을 연구하였다. SNS 메시지의 심리적 거리 분석을 위해 코더들을 활용하여 내용분석을 수행하였으며, 이와 같은 방법으로 레이블링된 데이터를 자연어 처리 방법을 이용하여 워드 임베딩을 수행함으로써 머신 러닝 수행을 위한 입력 데이터를 마련하였다. 머신 러닝 분석법 중 Support Vector Machine(SVM)을 이용하여 SNS 메시지와 심리적 거리 간의 관계를 학습시켰으며, 마지막으로 테스트 데이터를 이용하여 심리적 거리를 예측함으로써 머신 러닝 분석의 성과를 검증하였다. 심리적 거리 측정 방법론 수행 결과, 코더들의 내용분석 결과가 특정 값으로 편향되어 SVM 예측의 민감도와 정밀도가 낮은 결과가 도출되었다. 심리적 거리 응답 비율을 보정하고 코더들의 1차 내용분석 결과 중 답변이 일치한 데이터로 한정지어 머신 러닝을 실행한 결과 심리적 거리 예측의 정확도, 민감도, 특이도, 정밀도 모두 향상되어 심리적 거리가 70% 이상 예측되는 성과를 보였다. 본 연구는 SNS 메시지의 심리적 거리를 측정하는 방법을 제시함으로써 독자와의 심리적 거리를 제어 가능한 전략 요소로 활용 가능하게 할 것이라 기대된다.

키워드 : SNS 마케팅, 심리적 거리, 자연어 처리, 머신 러닝, 서포트 벡터 머신

Ⅰ. 서 론

무선 네트워크와 스마트폰이 발전함에 따라 소셜 네트워크 서비스(이하 SNS)는 사람들 간 소통의 주요 매체로 자리잡았다. SNS를 통해 서로의 근황을 공유함으로써 기존의 인간관계를 유지하고 자신의 관심사와 정보를 알림으로써 새로운 인간관계를 형성하는 것이 자연스러운 일이 되었다. SNS

† 이 논문은 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016S1A5A8020010).

는 인터넷 시대의 새로운 대화 수단이 된 것이다. 회사들 또한 SNS를 고객과 직접 소통할 수 있는 효과적인 커뮤니케이션 채널로 인식하고 있다. 포춘 500대 회사 중 98%, 88%, 85%의 회사들이 LinkedIn, Twitter, Facebook을 운영하는 등(Barnes and Pavao, 2018) 많은 회사들이 SNS를 이용하여 회사와 제품을 홍보하고 고객으로부터 의견을 청취하는 양방향 커뮤니케이션을 하고 있다(Burson-Marsteller, 2012).

면대면 커뮤니케이션에는 상대방의 생각과 감정을 이해하는데 도움이 되는 많은 단서들이 존재한다. 대화 내용뿐 아니라 억양, 표정, 몸짓 등이 모두 상대방을 이해하는 실마리가 된다(Rheingold, 1998). 그러나 컴퓨터 매개 커뮤니케이션 채널인 SNS에서는 이와 같은 실마리를 사용하는 것이 불가능하다(Rheingold, 1998). 대신 면대면 커뮤니케이션에서는 사용 불가한 이미지나 이모티콘과 같은 실마리를 이용하여 자신의 의견이나 감정을 간접적, 유희적으로 전달하며(강명수, 2005), 가상 청중을 상정하여 적절한 내용과 어휘, 어조 등을 사용하는 메시지 전략을 활용한다(Marwick and Boyd, 2010). 불특정 다수가 청자가 될 수 있는 만큼 SNS 메시지를 작성함에 있어서 말하고자 하는 의미와 감정을 효과적으로 전달하기 위한 전략을 이용하는 것이다.

SNS가 커뮤니케이션의 한 수단이 됨에 따라 SNS를 통해 주고 받는 메시지에 대한 연구들이 많이 행해지고 있다. 초기 연구들은 내용 분석 방법을 이용하여 SNS 메시지의 유형(McCorkindale, 2010; Naaman *et al.*, 2010; 이수범, 김남이, 2012; 이은선, 김미경, 2012), 메시지 토픽(Hum *et al.*, 2011; McCorkindale, 2010; Zhang *et al.*, 2010; 이수범, 김남이, 2012), 어조(이은선, 김미경, 2012) 등 메시지의 특징 등을 분석하였으며, 최근에는 자연어 처리 기법과 머신 러닝 방법을 적용하여 토픽 모델링(Al-garadi *et al.*, 2016; 배정환 등, 2014)과 감성분석(Cho *et al.*, 2014; Hemalatha *et al.*, 2013; Kaur and Saini, 2014; Wang *et al.*, 2013) 등을 수행함으로써 SNS 메시지 분석에 객관적인 분석 방법론을 연구하고 있다. 이와 같은 SNS 메시지 분석 연구들은 메시지의 주제나 긍정/부정과 같은 감성분석에 집중되고 있으며, SNS 메시지의 다양한 특징에 대한 연구는 부족한 실정이다.

SNS 메시지에 내포된 상대방과의 심리적 거리에 대한 내용분석 연구 역시 거의 수행되지 않았다. 커뮤니케이션 분야의 연구에 따르면 사람들이 SNS 메시지를 작성할 때 가상의 독자를 가정하고 메시지를 작성한다. 독자를 친구와 같이 가까운 관계로 가정하면 친근한 내용, 어휘와 어조를 사용하며, 거리감이 있는 공적인 청중을 독자로 가정하면 내용 면이나 형식면에서 격식을 갖춘 정중한 내용과 언어 스타일을 사용하게 될 것이다. 하지만, 아직까지 SNS 메시지에 포함된 글 작성자가 가정하고 있는 독자와의 심리적 거리를 분석하는 연구는 행해지지 않았다.

이에 본 논문에서는 SNS 메시지에 내포된 심리적 거리를 측정하는 방법을 연구함으로써 SNS 메시지 작성 전략에 있어서 심리적 거리를 고려할 수 있는 기반을 마련하고자 하였다. 화자와 청자 간 상호관계를 기반으로 하는 관계형 SNS인 페이스북을 대상으로 회사들의 팬페이지에 업로드된 메시지를 수집하고, 두 유형의 내용분석 방법을 순차적으로 활용하여 SNS 메시지에서 심리적 거리를 분석하였다. 첫 내용 분석 방법은 전통적인 내용 분석 방법으로, 코더들을 활용하여 메시지의 내용을 분석하는 방법(content analysis)이다. 이를 통해 코더들이 판단한 SNS 메시지에 내포된 심리적 거리 현황을 파악할 수 있었다. 그리고, 이 코딩 결과 데이터는 두 번째 내용 분석 방법 시행을 위한 투입 데이터를 제공하였다. 두 번째 분석 방법은 자연어 처리 기법과 머신 러닝 방법을 적용한 시스템적인 내용 분석 방법이다. 첫 분석 방법을 통해 레이블링(labeling)된 SNS 메시지를 머신 러닝의 투입(input) 데이터로 활용하여 심리적 거리를 학습하였으며 SNS 메시지의 심리적 거리를 예측하는 실험을 통해 그 성능을 분석함으로써 머신 러닝을 이용한 내용 분석 방법의 효과를 검증하였다.

본 연구는 이전 연구들이 살펴보지 않았던 SNS 메시지에 내포된 심리적 거리에 초점을 맞춰 이를 측정할 연구 방법을 제안했다는 점에서 이론적 의의를 찾을 수 있다. 전통적인 코더를 활용한 내용분석 방법에 자연어 처리 기법과 머신 러닝 방법을 추가적으로 적용함으로써 주관적인 심리적 거리 개념을 객관적이고 시스템적인 방법을 이용하여 측정하고자 하였다.

본 연구는 실무적으로도 다음과 같은 의의를 가진다. SNS 메시지를 작성함에 있어서 가정하고 있는 청중과 적절한 심리적 거리를 유지하고 있는지 판단할 수 있는 방법을 제시하였다. 이는 SNS 마케팅 전략에 있어서 적절한 소구층을 결정하는데 기초 자료로 활용될 수 있으며, 효과적인 SNS 메시지 작성을 위한 가이드를 제공해 줄 것이다.

Ⅱ. 관련 연구

2.1 SNS 메시지의 자연어 처리

이제까지 많은 연구들이 SNS 메시지의 특징을 분석해 왔다. 내용분석 방법을 이용하여 SNS 메시지의 유형(McCorkindale, 2010; Naaman *et al.*, 2010; 박종필, 손재열, 2012; 이수범, 김남이, 2012; 이은선, 김미경, 2012; 조태종 등, 2012)과 어조(이은선, 김미경, 2012), 그리고 메시지 내용(Hum *et al.*, 2011; McCorkindale, 2010; Zhang *et al.*, 2010; 이수범, 김남이, 2012)을 분석함으로써 효과적인 SNS 메시지와 다양한 특징의 SNS 메시지에 대한 사용자 태도 등을 연구해 왔다. 하지만 이러한 연구들이 사용한 코더들의 내용 분석 방법은 주관적 판단이 배제되기 어렵고 같은 콘텐츠에 대해서도 상황에 따라 다른 판단을 내릴 수 있다는 한계가 있다. 따라서, 최근에는 자연어 처리 방법(이하 NLP)과 나아가 인공지능 기법을 적용함으로써 SNS 메시지 분석 결과의 신뢰도를 높이고 있다(Al-garadi *et al.*, 2016; Cho *et al.*, 2014; Kaur and Saini, 2014; Lee *et al.*, 2012; 이태원, 홍태호, 2015).

NLP를 이용하여 SNS 메시지를 분석한 많은 연구들이 감성분석(sentiment analysis) 방법을 활용하였다. Wang *et al.*(2013)은 감성분석 방법을 이용하여 SNS 메시지에서 우울증을 탐지하는 모델을 개발하였으며, Hemalatha *et al.*(2013)과 김태환 등(2014)도 감성분석을 활용하여 SNS 메시지의 긍정/부정/중립적 성향을 구분하는 연구를 수행하였다. 배정환 등(2014)은 NLP 기법 중 토픽모델링 방법을 이용하여 트위터 상의 이슈를 분석하고 시각화하는 시스템을 제안하였다. 최근의 연구들은 NLP 기법에 인공지능 방법론을 추가적으로 적용함으로써 내용 분석 결과의 정확도를 높이고 있다. 지도 학습(supervised machine learning) 방법을 활용하여 트위터 메시지에서 사이버폭력 여부를 구분하고(Al-garadi *et al.*, 2016), SNS 메시지에 내포된 감정을 분석함에 있어서 Support Vector Machine(SVM)(Cho *et al.*, 2014; Kaur and Saini, 2014)나 Naive Bayes(Cho *et al.*, 2014; Eliaçık and Erdoğan, 2015; Lee *et al.*, 2012)를 이용하는 등 기계학습 방법을 이용하여 SNS 메시지에 대한 학습 결과를 기반으로 SNS 메시지의 특징 분류의 정확도를 높이는 분석 방법론들을 활용하고 있다.

본 논문에서는 회사 SNS를 대상으로 이제까지 SNS 메시지에서 분석되지 않은 심리적 거리를 NLP와 기계학습 방법을 통해 분류하고 검증함으로써 SNS 메시지의 중요한 특징 중 하나인 화자와 청자 간의 거리감을 객관적으로 분석한다.

2.2 심리적 거리

심리적 거리는 특정 대상이 나와 심리적으로 떨어져 있는 정도이다(Trope, 2004). 해석수준이론(construal level theory)에서는 심리적 거리를 4개의 하위 차원으로 나누고 있는데, 시간적, 공간적, 가설적, 사회적 거리감이 그것이다(Stephan *et al.*, 2010). 본 연구에서는 SNS 메시지에 내포된 화자와 청자 간의 거리, 즉 사회적 거리에 초점을 맞춰 심리적 거리를 분석하였다.

사회적 거리는 나 자신으로부터 어떤 대상이 얼마나 사회적으로 멀리 떨어져 있는지에 대한 주관적 느낌이다(김재휘 등, 2012). 사회적 거리는 직접 측정하기에 추상적인 개념이므로 많은 학자들이 대리 개념들을 이용하여 간접적으로 측정하였다. 가족, 친구, 이방인 등과 같은 사람들 간의 관계(Bar-Anan and Liberman, 2006; Bogardus, 1933; Hall, 1969)를 사회적 거리를 추정하는 기준으로 삼기도 하였으며, 친근감(Argyle and Dean, 1965; Kim and Kim, 2011; Stephan *et al.*, 2010, 2011), 유사성(Kim and Kim, 2011; Liviatan *et al.*, 2008; Stephan *et al.*, 2011), 인지된 상호작용 수준(Kim and Kim, 2011), 정중함의 정도(Stephan *et al.*, 2010)로 측정하기도 하였다.

본 연구에서는 회사 SNS에 내포된 심리적 거리의 수준을 화자와 청자 간의 가정된 사회적 관계, 친밀감 등 다양한 측정 변수를 이용하여 내용 분석하고, 그 데이터를 이용하여 자연어처리 기법과 기계학습 방법을 활용하여 학습하고 예측함으로써 심리적 거리를 측정하는 시스템적인 방법을 연구하였다.

Ⅲ. 연구 방법

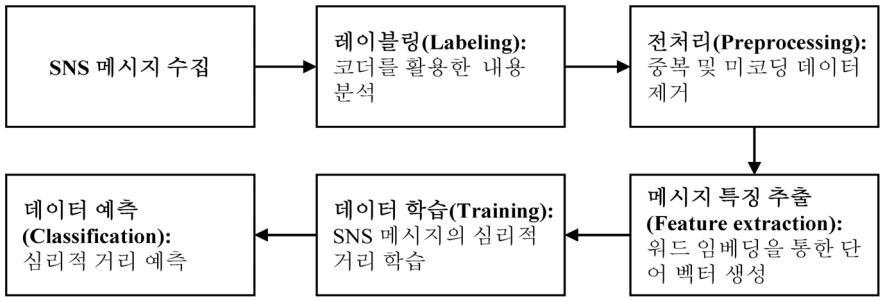
SNS 메시지에서 머신 러닝 방법을 이용하여 심리적 거리를 예측하는 본 연구는 <그림 1>과 같은 단계로 진행하였다. 페이스북의 회사 팬페이지를 선정하여 메시지를 크롤링하였으며, 코더들을 통해 해당 메시지를 내용 분석함으로써 SNS 메시지에 내포된 심리적 거리를 레이블링하였다. 중복 및 미코딩 데이터를 제거하는 데이터 전처리 과정을 거쳐 워드 임베딩을 통해 단어 벡터를 생성하고 Linear SVM을 이용하여 학습하고 심리적 거리를 예측하는 실험을 수행하였다. SNS 메시지를 전처리하고 머신 러닝을 이용하여 학습·예측하는 모듈은 Python을 이용하여 코딩하였다.

3.1 데이터 수집

본 연구에서는 SNS 사이트 중 페이스북을 대상으로 회사 팬페이지에 포스팅된 메시지를 분석하였다. 페이스북은 많은 회사들이 마케팅 목적으로 이용하고 있는 전 세계적으로 큰 SNS 사이트이며, 이용자를 중심의 관계 형성과 친목 유지가 목적인 관계형 SNS(안대천, 김상훈, 2012)이므로 상호 간의 심리적 거리가 중요한 SNS라는 특징을 갖는다.

페이스북 메시지는 Facebook Graph API를 이용하여 수집하였다. SNS 랭킹 사이트인 SocialBakers.com에서 한국 회사 페이스북 중 상위 50개 페이스북을 대상으로 2017년 7월부터 11월까지(5개월간) 총 7,517개의 메시지를 수집하였다.

SNS 메시지에 내포된 심리적 거리에 대한 레이블링(labeling) 작업은 코더들을 통한 내용분석(content analysis) 방법을 이용하였다. 커뮤니케이션 행동은 메시지의 내용(말하고자 하는 내용)과 메시지의 형식(어휘, 어투, 언어학적 스타일)의 두



〈그림 1〉 연구 방법 진행 단계

차원에서 살펴봐야 한다는 Altman(1975)의 주장에 맞춰 SNS 메시지의 심리적 거리를 내용 차원에서 2개 항목(친근감 여부, 개인내용 포함여부), 형식 차원에서 4개의 항목(정중함 여부, 완성문장 여부, 줄임말 사용여부, 이모티콘 사용여부)으로 코딩 스킴을 작성하였다(<부록> 참고). 친근감 여부와 개인정보 포함여부는 마케팅 광고의 효과에 영향을 미치는 사회적 유대(social tile)의 구성 요소이다(Adler and Kwon, 2002; Nahapiet and Ghoshal, 1998; Shen *et al.*, 2016; Wellman and Berkowitz, 1998). 그리고 정중함 여부, 완성문장 여부, 줄임말 사용여부, 이모티콘 사용여부는 친근감에 영향을 미치는 특징(feature)들이다. 친근감, 가까움으로 대변되는 사회적 유대가 높으면 광고에 대한 구전(word-of-mouth) 동기가 높아지며 광고에 대한 긍정적인 태도가 형성된다(Shen *et al.*, 2016; Wellman and Berkowitz, 1998). 광고의 효과에 있어서 사회적 유대감이 중요한데, 친근감 등의 심리적 거리가 이에 영향을 미치는 것이다. 따라서, 본 연구에서 분석하고자 하는 심리적 거리의 하위 변수이면서 마케팅의 영향인자인 사회적 유대의 하위변수인 친근감, 개인정보 오픈 여부, 그리고 이에 영향을 미치는 특성 항목들을 코딩 항목으로 설정하였다.

내용분석은 다음과 같은 절차로 진행하였다. 총 14명의 대학생 코더들이 2명씩 팀을 이루어 동일한 메시지에 대해 각각 심리적 거리를 코딩한 후 일치하지 않는 결과에 대해서는 다른 팀의 코더가 확정 짓는 방식으로 내용분석을 시행함으로써 머신 러닝으로 학습하고 예측할 데이터셋을 완성하였다. 주관적 개념인 심리적 거리를 동일 관점에서 객관적으로 코딩하도록 하기 위하여 코더들을 대상으로 코딩 설명회를 개최하였다. 이 설명회를 통해 코딩 스킴을 설명하였으며, 실제 수집한 SNS 메시지 중 20개의 선별하여 함께 코딩해 봄으로써 코딩 스킴에 대해 논의하고 보완하는 작업을 수행하였다. '완성문장 여부', '줄임말 사용여부' 등에서 불완전한 문장이나 줄임말이 1개라도 포함되면 '아니오'로 코딩하는 등 코딩 규칙을 상세 정의함으로써 코더들 간 일관된 코딩 기준을 정립하였다.

심리적 거리 하위 항목별 코딩 결과 데이터 건수(코딩되지 않은 데이터 건수 제외)는 <표 1>과 같다. 친근감 여부 항목은 '예'로 코딩한 비율이 54.89%, '아니오'로 코딩한 비율이 45.11%로 유사하였으나, 그 외 항목들은 80% 이상의 메시지가 '아니오'로 코딩됨으로써 데이터 편중이 심한 것으로 나타났다.

3.2 데이터 전처리

레이블링한 데이터를 사용하기 전, 데이터 무결성을 높이기 위하여 SNS 메시지가 중복되는 데이터와 레이블링 과정에서 항목별 코딩되지 않은 데이터를 제거하는 전처리 작업을 수행하였다. 제거된 데이터 건수는 총 73개였으며, 이를 제외한 총 7,444개의 데이터를 대상으로 분석을 계속 시행하였다.

〈표 1〉 회사 SNS 메시지 코딩 결과

항목	답변	건수	비율(%)	항목	답변	건수	비율(%)
친근감 여부	예	4,120	54.89	완성문장 여부	예	1,045	13.92
	아니오	3,386	45.11		아니오	6,460	86.08
개인내용 포함여부	예	251	3.34	줄임말 사용여부	예	1,191	15.85
	아니오	7,261	96.66		아니오	6,321	84.15
정중함 여부	예	1,443	19.21	이모티콘 사용여부	예	1,093	14.54
	아니오	6,068	80.79		아니오	6,423	85.46

다음으로는 SNS 메시지에 대한 형태소 분석을 실시하였다. 처음에는 오픈소스인 KoNLPy를 이용하여 형태소 분석을 시행하였으나, 하나의 고유명사가 두 개의 형태소로 분리되거나 외국어가 포함된 경우 오류를 반환하는 등의 문제점이 발생하였다. 이에 한국전자통신연구원(http://aiopen.etri.re.kr)이 제공하고 있는 '언어분석 API'를 이용하여 형태소를 분석하였다. 분석 결과 분류된 품사들을 모두 이용하여 이후 분석 프로세스를 진행하였다.

3.3 텍스트 피쳐 추출

SNS 메시지의 심리적 거리를 머신 러닝을 이용하여 학습하고 예측하기 위해서는 SNS 메시지를 구성하는 형태소를 이용하여 수치화된 단어 벡터(집합)를 만들어야 한다(Tripathy *et al.*, 2015). Bag-of-Words (BOW) 모델은 텍스트를 단어의 집합으로 표현하는 방법 중 하나로, 문장 안에 포함된 단어와 단어의 개수나 단어의 중요도 등 관련 수치를 재정렬한 벡터이다(Goldberg, 2017). BOW 모델에서는 단어의 순서나 문장 구조는 사라지지만, SNS 메시지에서 느껴지는 심리적 거리는 문장 구조보다는 사용하는 단어와 연관성이 높다고 판단하여 BOW 모델을 이용하였다.

본 연구에서는 BOW를 만드는 방법으로 CountVectorizer와 TfidfVectorizer를 이용하였으며, Python 머신 러닝 패키지인 Scikit-learn을 이용하여 분석하였다. CountVectorizer는 문장을 구성하는 단어들의 빈도를 카운트하는 방식으로 벡터 행렬을 만든다(Tripathy *et al.*, 2015). 반면, TfidfVectorizer는 단어들의 중요도를 기반으로 가중치를 계산하는 TF-IDF (Term Frequency-Inverse Document Frequency) 방식을 적용하여 벡터 행렬을 만든다. 이를 통해 '은', '는'과 같은 조사나 여러 SNS 메시지에서 반복 사용되는 중요도가 떨어지는 단어들의 가중치를 낮추는 방식으로 벡터 행렬을 만들 수 있다(Tripathy *et al.*, 2015).

3.4 머신 러닝

본 연구에서는 머신 러닝 방법을 이용하여 SNS 메시지의 심리적 거리를 학습시키고 분류하였다. 기계 스스로 데이터를 학습하여 유의미한 정보를 도출하게 함으로써 SNS 메시지를 내용 분석함에 있어서 문제점이라 할 수 있는 주관적 판단을 배제하도록 하였다.

본 연구에서는 머신 러닝의 클러스터링 기법 중 하나인 Support Vector Machine(SVM)을 이용하였다. SVM은 서로 다른 두 집단을 분류하는 초평면(hyperplane)들 중 두 집단 사이의 거리를 최대로 하는 초평면을 찾는 알고리즘이다(Cortes and Vapnik, 1995; Tripathy *et al.*, 2015). SVM은 일반화의 오류가 낮고 계산 비용이 적게 든다는 장점이 있다.

SVM의 초평면은 식 (1)로 표현된다.

$$w_0 \cdot z + b_0 = 0 \tag{1}$$

여기서 z는 분류하고자 하는 데이터, w_0는 초평면에 최적인 가중치로서 초평면의 법선 벡터를 나타내며 식 (2)와 같이 표현할 수 있다.

$$w_0 = \sum_{support\ vector} \alpha_i z_i \tag{2}$$

초평면을 기준으로 식 (3)과 같이 데이터에 대해 +1과, -1로 선형 분리하여 값을 찾아 분류해 내는 것이 SVM의 원리이다.

$$\begin{aligned} w \cdot x_i + b \geq 1 \quad if \quad y_i = 1 \\ w \cdot x_i + b \leq -1 \quad if \quad y_i = -1 \end{aligned} \tag{3}$$

Ⅳ. 실험 및 결과

4.1 실험 방법

본 연구에서는 회사별로 정렬되어 있는 SNS 메시지를 무작위로 섞은 후, 전체 데이터의 80%에 해당하는 5,955건을 SNS 메시지의 심리적 거리를

학습하는 데이터로 이용하였으며 20%에 해당하는 1,489건을 이용하여 테스트를 진행하였다.

CountVectorizer와 TfidfVectorizer를 이용하여 단어 벡터를 만들었으며, 단어의 최소 출현 횟수(min_df)를 각각 1과 2로 설정하여 총 4개 유형의 벡터를 만들었다. SNS 메시지의 심리적 거리 학습은 linear SVM을 이용하여 시행하였다.

4.2 실험 결과

SVM을 이용하여 학습하고 1,489건의 데이터를 이용하여 SNS 메시지의 심리적 거리를 예측하였다. 심리적 거리 하위 변수들에 대한 여부 답변을 맞춘 결과는 <표 2>와 같다. TP, TN은 각각 true positive, true negative의 약자로, '예'라는 정답을 '예'라고 예측하고 '아니오'라는 정답을 '아니오'라고 예측한 횟수이며, FP와 FN은 false positive와 false negative의 약자로 실제 값과 예측 값이 다른 잘못 예측한 횟수이다. <표 3>은 이러한 예측의 정답 유무 횟수를 기반으로 정확도(accuracy), 민감도(sensitivity), 특이도(specificity), 정밀도(precision)를 계산하여 SVM 성과를 분석한 결과이다. 심리적 거리의 하위 변수 중 친근감 여부에 대해서는 모든 예측 성능이 0.70 이상으로 양호했으나, 그 외의 변수들에 대해서는 민감도와 정밀도가 현저히 낮은 결과를 보였다. 워드 임베딩 방법에 있어서는 TfidfVectorizer를 이용한 경우의 성능이 더 나은 것으로 분석되었다. 즉, 단어의 중요도에 따라 가중치를 사용하는 것이 SVM의 성능을 향상시키는 것으로 밝혀졌다. 하지만, min_df, 즉 단어의 최소 출현 회수는 SVM을 이용한 SNS 메시지의 심리적 거리 예측에 있어서 유의미한 영향을 미치지 않았다.

〈표 2〉 심리적 거리의 하위 변수 예측 적중 여부 결과

분석 방법	구 분	친근감 여부	개인내용 포함여부	정중함 여부	완성문장 여부	줄임말 사용여부	이모티콘 사용여부
CountVectorizer, min_df=1	TP	621	5	207	123	78	83
	FP	182	15	75	79	92	90
	FN	178	37	105	76	150	146
	TN	508	1,432	1,102	1,211	1,169	1,170
CountVectorizer, min_df=2	TP	614	6	209	120	80	90
	FP	192	32	86	89	128	113
	FN	185	36	103	79	148	139
	TN	498	1,415	1,091	1,201	1,133	1,147
TfidfVectorizer, min_df=1	TP	654	0	184	85	37	43
	FP	186	0	45	16	23	17
	FN	145	42	128	114	191	186
	TN	504	1,447	1,132	1,274	1,238	1,243
TfidfVectorizer, min_df=2	TP	654	0	188	85	45	47
	FP	189	0	45	21	25	20
	FN	145	42	124	114	183	182
	TN	501	1,447	1,132	1,269	1,236	1,240

* TP(true positive): 실제 참인데 예측도 참, FP(false positive): 실제 거짓인데 예측은 참, FN(false negative): 실제 참인데 예측은 거짓, TN(true negative): 실제 거짓인데 예측도 거짓.

〈표 3〉 SVM 분석 성과 결과

분석 방법	측정치	친근감 여부	개인내용 포함여부	정중함 여부	완성문장 여부	줄임말 사용여부	이모티콘 사용여부
CountVectorizer, min_df=1	정확도	0.76	0.97	0.88	0.90	0.84	0.84
	민감도	0.78	0.12	0.66	0.62	0.34	0.36
	특이도	0.74	0.99	0.94	0.94	0.93	0.93
	정밀도	0.77	0.25	0.73	0.61	0.46	0.48
CountVectorizer, min_df=2	정확도	0.75	0.95	0.87	0.89	0.81	0.83
	민감도	0.77	0.14	0.67	0.60	0.35	0.39
	특이도	0.72	0.98	0.93	0.93	0.90	0.91
	정밀도	0.76	0.16	0.71	0.57	0.38	0.44
TfidfVectorizer, min_df=1	정확도	0.78	0.97	0.88	0.91	0.86	0.86
	민감도	0.82	0	0.59	0.43	0.16	0.19
	특이도	0.73	1	0.96	0.99	0.98	0.99
	정밀도	0.78	NA	0.80	0.84	0.62	0.72
TfidfVectorizer, min_df=2	정확도	0.78	0.97	0.89	0.91	0.86	0.86
	민감도	0.82	0	0.60	0.43	0.20	0.21
	특이도	0.73	1	0.96	0.98	0.98	0.98
	정밀도	0.78	NA	0.81	0.80	0.64	0.70

심리적 거리를 코딩한 결과 데이터가 특정 값에 편향된 점을 고려하여, 가장 성능이 좋았던 TfidfVectorizer 방법으로 벡터 행렬을 생성하고 단어를 1번만 이용하여 SVM을 시행한 예측 상세 결과를 정리하였다. 심리적 거리 하위 항목의 응답별로 일치 답변 개수와 일치율을 정리하였는데, 그 내용은 <표 4>와 같았다. 각 심리적 거리 항목에 대한 전체 응답의 일치율은 모두 77% 이상으로 높게 나타났지만, 각 응답에 대한 일치율은 좋지 않았다. 특정 값으로 치우친 항목에 대한 머신 러닝 예측 결과, 빈도가 높은 편중된 답변으로 예측하는 경향이 뚜렷하게 나타났다. 답변율(예 = 54.89%, 아니오 = 45.11%)이 비슷했던 '친근감 여부' 항목에 대한 예측율에 대해서는 각각 81.85%, 73.04%로 잘 예측한 것으로 분석되었다.

특정 값으로 편향된 응답의 분포가 머신 러닝 분석 결과의 신뢰도를 저하시키는 문제를 해결하기 위하여 응답의 비율을 동일하게 보정하여 SVM을 재실행하는 추가 분석을 시행하였다. 응답 비중이 비슷했던 친근감 여부 항목과 과도하게 편향되어 데이터 비율 보정이 어려운 메시지 대상, 개인내용 포함여부 항목을 제외한 4개의 항목(정중함 여부, 완성문장 여부, 줄임말 사용여부, 이모티콘 사용여부)에 대해 예/아니오 응답을 랜덤하게 추출하여 응답별로 1,000개의 데이터셋, 총 2,000개의 학습 데이터를 만들어 학습시킴으로써 데이터 편향의 문제를 해결하였다. 학습 결과를 기반으로 실행한 심리적 거리 예측에는 예/아니오 응답 비율이 동일한 400개의 테스트 데이터를 이용하였다. 이와 같이 응답의 분포를 보정한 후 SVM을 실행한 결과는 <표 5>와 같다. 정확도, 민감도, 특이도, 정밀도 모두 0.78 이상으로, 심리적 거리 예측 성능이 현저하게 향상된 것으로 나타났다.

SNS 메시지에 내포된 심리적 거리는 주관적인 개념으로, 본 연구에서 코더들을 활용한 첫 내용 분석 코딩 결과의 코더간 일치율이 지나치게 낮았었다. 글을 읽고 느끼는 글쓴이와의 거리감의 기준은 사람들마다 다르며, 따라서 심리적 거리의 수준을 두 사람이 동일하게 판단하기 어렵다. 이에 본 연구에서는 한 팀으로 구성된 2명의 코더가

1차 내용 분석에서 동일하게 코딩한 데이터만을 이용하여 심리적 거리 수준을 학습하고 예측한 두 번째 추가 분석을 시행하였다. 메시지 대상 항목에 대해서는 특정 응답에 코딩 결과값의 몰림 현상이 두드러져 제외하였으며, 친근감 여부, 정중함 여부, 완성문장 여부, 줄임말 사용 여부, 이모티콘 사용 여부 항목에 대해서만 추가 분석하였다. 1차 내용 분석에서 동일하게 답변된 데이터만을 추출하고 '예', '아니오' 응답 비율을 1:1로 맞춰 코딩 결과 수가 많았던 데이터를 무작위로 선별함으로써 특정 답변 데이터가 더 많이 학습되는 경우의 수를 제거하였다. 이와 같이 선별된 데이터의 80%는 학습을 위해 사용하였으며, 나머지 20%의 데이터로 심리적 거리를 예측하였다. <표 6>은 추가 분석에 이용한 데이터 개요이며, <표 7>은 추가 분석한 SVM을 활용한 심리적 거리 예측 결과이다. 줄임말 사용여부와 이모티콘 사용여부 예측 결과가 이전 추가 분석의 결과보다 좋지 않았다. 하지만 모든 SVM 예측 성과 지표가 0.70을 넘었으며, 이전 추가 분석에 활용한 데이터 건수보다 약 1,000건 적었던 점을 고려하면 SVM이 심리적 거리를 잘 예측했다 할 수 있다.

〈표 4〉 원본 데이터와 SVM(TfidVectorizer, min_df=1) 실험 데이터 비교

항목	응답 유형	실제 응답	일치 응답	일치율(%)
친근감 여부	예	799	654	81.85
	아니오	690	504	73.04
	전체	**1,489**	**1,158**	**77.77**
개인내용 포함여부	예	42	0	0
	아니오	1,447	1,447	100.00
	전체	**1,489**	**1,447**	**97.18**
정중함 여부	예	312	184	58.97
	아니오	1,177	1,132	96.18
	전체	**1,489**	**1,316**	**88.38**
완성문장 여부	예	199	85	42.71
	아니오	1,290	1,274	98.76
	전체	**1,489**	**1,359**	**91.27**
줄임말 사용여부	예	228	37	16.23
	아니오	1,261	1,238	98.18
	전체	**1,489**	**1,275**	**85.63**
이모티콘 사용여부	예	229	43	18.78
	아니오	1,260	1,243	98.65
	전체	**1,489**	**1,286**	**86.37**

〈표 5〉〔추가분석 1〕 SVM 실행 결과(TfidfVectorizer, min_df=1)

측정치	정중함 여부	완성문장 여부	줄임말 사용여부	이모티콘 사용여부
정확도	0.90	0.89	0.81	0.82
민감도	0.88	0.93	0.83	0.78
특이도	0.91	0.85	0.79	0.85
정밀도	0.91	0.86	0.80	0.84

〈표 6〉〔추가분석 2〕 분석 데이터 개요

구 분		친근감 여부	정중함 여부	완성문장 여부	줄임말 사용여부	이모티콘 사용여부
1차 코딩 결과 데이터	예	1,994	722	661	481	813
	아니오	2,179	5,259	5,612	5,352	6,189
분석 대상 총 건수		3,988	1,444	1,322	962	1,626
학습 건수		3,190	1,155	1,057	769	1,300
테스트 건수		798	289	265	193	326

〈표 7〉〔추가분석 2〕 SVM 예측 결과(TfidfVectorizer, min_df=1)

분석 구분	친근감 여부	정중함 여부	완성문장 여부	줄임말 사용여부	이모티콘 사용여부
정확도	0.88	0.91	0.90	0.73	0.75
민감도	0.86	0.88	0.86	0.70	0.73
특이도	0.90	0.94	0.95	0.75	0.76
정밀도	0.89	0.94	0.94	0.75	0.76

Ⅴ. 결 론

본 논문에서는 SNS 메시지에 내포된 심리적 거리에 초점을 맞춰 이를 측정하기 위한 방법에 대해 연구하였다. 분석 대상 메시지로 회사들의 페이스북 팬페이지에 게시된 메시지를 수집하였으며, 심리적 거리는 두 가지 분석 방법을 통하여 추출하였다. 첫 번째 방법으로 코더들을 이용하여 심리적 거리를 추출하는 내용분석을 수행하였는데, 회사들이 SNS 메시지를 작성함에 있어서 가상 청중을 공적인 관계로 가정하고 있었으며 그 코딩 결과가 특정 값에 지나치게 편향됨을 알 수 있었다. 이 내용분석 결과 데이터는 두 번째 내용 분석 방법인 SVM 머신 러닝 분석의 학습 및 테스트 데이터로 활용되었다. 편향된 데이터 분포 때문에 심리적 거리 예측율의 신뢰도가 떨어졌으나, 응답 값의 비율을 1:1로 보정한 후 실행한 SVM 분석 결과는 양호하고 안정적인 예측율을 보여주었다. 본 연구는 SNS 메시지에 내포된 주관적인 심리적 거리를 자연어 처리 기법과 머신 러닝 방법을 이용하여 체계적으로 측정할 수 있음을 검증하였다.

본 연구를 수행함에 있어서 몇 가지 한계점들이 있었다. 첫째, 심리적 거리의 정도를 정량화하는 과정에서 주관적 판단을 완전히 배제하기 어려웠다. SNS 메시지를 읽고 느끼는 글쓴이와의 친근감과 공감은 평소 개인들이 가지고 있는 인간관계 유형과 가치관에 따라 다를 수밖에 없으므로(이희경, 2001), 두 명의 코더가 동일 메시지에 대해 분석한 코딩 작업 결과의 일치율이 낮았다. SNS 메시지의 심리적 거리가 정확하게 레이블링 된다면 머신 러닝 분석 결과의 신뢰도 또한 높아질 것이다. 둘째, SNS 메시지에 대한 자연어 처리 방법에 있어서 단어의 개수(CountVectorizer)와 단어 중요도(TfidfVectorizer)를 기반으로 단어 벡터를 생성하는 방법만을 활용하였다. 심리적 거리 단어 사전(dictionay) 이용 등 추가적인 자연어 처리 기법을 적용한다면 심리적 거리 측정의 정확도를 높일 수 있을 것이라 생각한다. 마지막으로, 본 연구에서는 머신 러닝 분석 방법 중 SVM 만을 적용하였다는 한계가 있다. SVM 외 다양한 머신 러닝 알고리즘을 적용하고 그 성능을 비교한다면 더 효과적인 심리적 거리 측정 방법을 찾을 수 있을 것이라 생각한다.

본 연구는 기술한 바와 같은 한계점들을 가지고 있지만, 이론적으로나 실무적으로 큰 의의를 가지고 있다. 본 연구는 이전 연구들이 살펴보지 않았던 SNS 메시지의 심리적 거리에 초점을 맞춘 연구이다. SNS 메시지 작성에 있어서 가정된 청중이 중요함에도 불구하고 연구 방법의 어려움 때문에 이에 대한 실증 연구가 부족했었다. 본 연구는 심리적 거리를 측정하는 방법을 제안하고 검증했다는 이론적 의의를 가진다. 둘째, 본 연구는 SNS 메시지의 내용을 두 가지 연구 방법을 단계적으로 이용하여 분석하였다. 코더를 활용한 전통적인 내용분석 방법과 자연어 처리 및 머신 러닝 분석 방법을 같이 이용하여 SNS 메시지의 심리적 거리를 레이블링하고 이를 체계적이고 객관적인 시스템 활용 내용 분석 방법으로 확장하였다. SNS 메시지 내용분석에 있어서 가장 큰 문제점인 주관적 판단의 개입을 해결할 방안을 검증했다는 의의를 가진다.

본 연구는 실무적으로도 큰 의의를 가지고 있다. SNS 마케팅 전략 수립에 참고할 심리적 거리 개념을 제안하였다. 회사와의 유대감이 광고에 대한 구전 동기 및 태도에 긍정적인 영향을 미치며, 이에 회사 SNS 메시지에 내포된 적절한 거리감은 브랜드 이미지와 태도에 영향을 미치는 영향 요인이다. 본 연구는 심리적 거리를 측정할 수 있는 방법을 제시함으로써 청자와의 적절한 심리적 거리를 제어 가능한 전략 요소로 활용 가능하게 하였다. 이는 SNS 마케팅 전략 수립에 있어서 적절한 소구층을 결정하는데 통찰력을 제공할 것이며, 효과적인 SNS 메시지 작성에 대한 가이드를 제공해 줄 것이다.

참 고 문 헌

[1] 강명수, “온라인 커뮤니티 형성과 유지에 관한 연구: 규범적 몰입과 감성적 몰입의 매개역할을 중심으로”, *대한경영학회지*, 제18권, 제1호, 2005, pp. 67-87.

[2] 김재휘, 김희연, 부수현, “소셜 미디어를 활용한 공공캠페인 커뮤니케이션 전략: 해석수준 이론에 따른 메시지 구성과 미디어에 대한 사회적 거리를 중심으로”, *광고학연구*, 제23권, 제1호, 2012, pp. 183-205.

[3] 김태환, 정우진, 이상용, “기업의 SNS 노출과 주식 수익률간의 관계 분석”, *Asia Pacific Journal of Information Systems*, 제24권, 제2호, 2014, pp. 233-253.

[4] 박종필, 손재열, “B2C 마이크로블로깅을 통한 고객참여 메커니즘의 이해”, *Asia Pacific Journal of Information Systems*, 제22권, 제4호, 2012, pp. 51-73.

[5] 배정환, 한남기, 송민, “토픽 모델링을 이용한 트위터 이슈 트래킹 시스템”, *Journal of Intelligence and Information Systems*, 제20호, 제2권, 2014. pp. 109-122.

[6] 안대천, 김상훈, “SNS 유형별 광고속성 평가 및 태도에 관한 연구: 블로그, 트위터, 페이스북, 유튜브의 비교”, *광고학연구*, 제23권, 제3호, 2012, pp. 53-84.

[7] 이수범, 김남이, “페이스북 팬페이지의 메시지 및 크리에이티브 전략에 관한 연구”, *소비자문제연구*, 제42권, 2012, pp. 123-148.

[8] 이은선, 김미경, “마케팅 커뮤니케이션 수단으로서의 기업 페이스북 팬페이지 이용행태 분석”, *광고학연구*, 제23권, 제2호, 2012, pp. 31-55.

[9] 이태원, 홍태호, “Support Vector Machine을 이용한 온라인 리뷰의 용어기반 감성분류모형”, *Information Systems Review*, 제17권, 제1호, 2015, pp. 49-64.

[10] 이희경, “공감수준과 친소관계가 따돌림에 대한 심리적 반응에 미치는 효과”, *教育心理硏究*, 제15권, 제3호, 2001, pp. 281-297.

[11] 조태종, 윤혜정, 이중정, “기업의 홍보 마케팅용 트위터의 리트윗 현황 분석: 이용자 특성과 콘텐츠 속성을 중심으로”, *Information Systems*

Review, 제14권, 제1호, 2012, pp. 21-35.

[12] Adler, P. S. and S. W. Kwon, "Social capital: Prospects for a new concept", *The Academy of Management Review*, Vol.27, No.1, 2002, pp. 17-40.

[13] Al-garadi, M. A., K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network", *Computers in Human Behavior*, Vol.63, 2016, pp. 433-443.

[14] Altman, I., *The Environment and Social Behavior: Privacy, Personal Space, Territory, Crowding*, Brooks/Cole Publishing Company, Monterey, California, 1975.

[15] Argyle, M. and J. Dean, "Eye-contact, distance and affiliation", *Sociometry*, Vol.28, No.3, 1965, pp. 289-304.

[16] Bar-Anan, Y. and N. Liberman, "The association between psychological distance and construal level: Evidence from an Implicit Association Test", *Journal of Experimental Psychology: General*, Vol.135, No.4, 2006, pp. 609-622.

[17] Barnes, N. G. and S. Pavao, "The 2017 Fortune 500 go visual and increase use of Instagram, Snapchat, and Youtube", 2018, Available at https://www.umassd.edu/cmr/socialmediaresearch/2017fortune500/#d.en.963986.

[18] Bogardus, E. S., "A social distance scale", *Sociology & Social Research*, Vol.17, 1933, pp. 265-271.

[19] Burson-Marsteller, "Global Social Media Check-up", 2012, Available at http://bm.com/social.

[20] Cho, S. W., M. S. Cha, S. Y. Kim, J. C. Son, and K. Sohn, "Investigating temporal and spatial trends of brand images using Twitter opinion mining", *2014 International Conference on Information Science & Applications (ICISA)*, 2014.

[21] Cortes, C. and V. Vapnik, "Support-vector networks", Machine learning, Vol. 20, 1995, pp. 273-297.

[22] Eliaçık, A. B. and N. Erdoğan, "User-weighted sentiment analysis for financial community on Twitter", *2015 11th International Conference on Innovations Information Technology(IIT)*, 2015, pp. 46-51.

[23] Goldberg, Y., *Neural Network Methods for Natural Language Processing*, Morgan & Claypool Publishers, 2017.

[24] Hall, E. T., *The Hidden Dimension*, Anchor Books, Doubleday & Company, inc., Garden City, NY., 1969.

[25] Hemalatha, I., G. P. S. Varma, and A. Govardhan, "Sentiment analysis tool using machine learning algorithms", *International Journal of Emerging Trends & Technology in Computer Science*, Vol.2, No.2, 2013, pp. 105-109.

[26] Hum, N. J., P. E. Chamberlin, B. L. Hambright, A. C. Portwood, A. C. Schat, and J. L. Bevan, "A picture is worth a thousand words: a content analysis of Facebook profile photographs", *Computers in Human Behavior*, Vol.27, 2011, pp. 1828-1833.

[27] Kaur, J. and J. R. Saini, "Emotion detection and sentiment analysis in text corpus: A differential study with informal and formal writing styles", *International Journal of Computer Applications*, Vol.101, No.9, 2014, pp. 1-9.

[28] Kim, B. K. and K. H. Kim, "The impact that social distance perceived by SNS affects communication", *ADADA2011 Kitakyushu*, 2011, pp. 201-202.

[29] Lee, H., Y. S. Choi, S. Lee, and I. P. Park, "Towards unobtrusive emotion recognition for affective social communication", *The 9th Annual IEEE Consumer Communications and Networ-*

king Conference, 2012, pp. 260-264.

[30] Liviatan, I., Y. Trope, and N. Liberman, "Interpersonal similarity as a social distance dimension: Implications for perception of others' actions", *Journal of Experimental Social Psychology*, Vol.44, 2008, pp. 1256-1269.

[31] Marwick, A. E. and D. Boyd, "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience", *New Media & Society*, Vol.13, No.1, 2010, pp. 114-133.

[32] McCorkindale, T., "Can you see the writing on my wall? A content analysis of the Fortune 50's Facebook social networking sites", *Public Relations Journal*, Vol.4, No.3, 2010.

[33] Naaman, M., J. Boase, and C. Lai, "Is it really about me? Message content in social awareness streams", *The Proceedings of CMC*, 2010, pp. 189-192.

[34] Nahapiet, J. and S. Ghoshal, "Social capital, intellectual capital and the organizational advantage", *Academy of Management Review*, Vol.23, 1998, pp. 242-266.

[35] Rheingold, H., "Virtual communities", in Frances Hesselbein et al.(eds.), *Academy The community of the future*, New York: The Drucker Foundation, 1998, pp. 115-122.

[36] Shen, G. C. C., J. S. Chiou, C. H. Hsiao, C. H. Wang, and H. N. Li, "Effective marketing communication via social networking site: The moderating role of the social tie", *Journal of Business Research*, Vol.69, No.6, 2016, pp. 2265-2270.

[37] Stephan, E., N. Liberman, and Y. Trope, "Politeness and psychological distance: A construal level perspective", *Journal of Personality and Social Psychology*, Vol.98, No.2, 2010, pp. 268-280.

[38] Stephan, E., N. Liberman, and Y. Trope, "The effects of time perspective and level of construal on social distance", *Journal of Experimental Social Psychology*, Vol.47, 2011, pp. 397-402.

[39] Tripath, A., A. Agrawal, and S. K. Rath, "Classification of sentimental reviews using machine learning techniquest", *Procedia Computer Science*, Vol.57, 2015, pp. 821-829.

[40] Trope, Y., "Theory in social psychology: Seeing the forest and the trees", *Personality and Social Psychology Review*, Vol.8, No.2, 2004, pp. 193-200.

[41] Wang, X., C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, "A depression detection model based on sentiment analysis in micro-blog social network", *The Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2013, pp. 201-213.

[42] Wellman, B. and S. D. Berkowitz, *Social Structures: A Network Approach*, Cambridge University Press, 1998.

[43] Zhang, J., Y. Sung, and W. N. Lee, "To play or not to play: An exploratory content analysis of branded entertainment in Facebook", *American Journal of Business*, Vol.25, No.1, 2010, pp. 53-64.

〈부록〉 SNS 메시지에 내포된 심리적 거리 코딩 스킴

구 분	변수	No.	스케일	설명
메시지의 심리적 거리	친근감 여부	메시지에서 친근감을 느끼셨습니까?		
		1	예	메시지를 읽고 작성자와 가까워진 느낌이 드는 경우
		2	아니오	메시지를 읽고 작성자와 가까워진 느낌이 들지 않는 경우
	개인내용 포함 여부	메시지가 작성자의 개인적인 내용을 포함하고 있습니까?		
		1	예	작성자 개인의 생각, 경험 등 개인적인 내용이 포함되어 있는 경우
		2	아니오	작성자 개인의 내용이 포함되어 있지 않은 경우
	정중함 여부	메시지의 언어 스타일이 정중하고 예의 바릅니까?		
		1	예	격식을 갖춘 공식적이고 정중하게 쓰여진 경우
		2	아니오	격식에 얽매이지 않고 일상적이며 허물없이 쓰여진 경우
	완성문장 여부	메시지가 완성된 문장을 사용하고 있습니까?		
		1	예	'다', '요' 등으로 끝남으로써 문장이 완성형 문장인 경우 모든 문장이 완성형일 때, '예'
		2	아니오	문장 끝이 술어로 끝나지 않은 비완성형 불완전 문장인 경우
	속어 사용 여부	메시지에 줄임말이 포함되어 있습니까?		
		1	예	줄임말이 사용된 경우
		2	아니오	줄임말이 사용되지 않은 경우
	이모티콘 사용 여부	메시지에 이모티콘이 포함되어 있습니까?		
		1	예	이모티콘이 사용된 경우 페이스북 자체 그림 이모티콘(크롤링 시 깨져보임) 제외
		2	아니오	이모티콘이 사용되지 않은 경우

Information Systems Review
Volume 21 Number 1
February 2019

A Study on the Extraction of Psychological Distance Embedded in Company's SNS Messages Using Machine Learning

Seongwon Lee* · Jin Hyuk Kim**

Abstract

The social network service (SNS) is one of the important marketing channels, so many companies actively exploit SNSs by posting SNS messages with appropriate content and style for their customers. In this paper, we focused on the psychological distances embedded in the SNS messages and developed a method to measure the psychological distance in SNS message by mixing a traditional content analysis, natural language processing (NLP), and machine learning. Through a traditional content analysis by human coding, the psychological distance was extracted from the SNS message, and these coding results were used for input data for NLP and machine learning. With NLP, word embedding was executed and Bag of Word was created. The Support Vector Machine, one of machine learning techniques was performed to train and test the psychological distance in SNS message. As a result, sensitivity and precision of SVM prediction were significantly low because of the extreme skewness of dataset. We improved the performance of SVM by balancing the ratio of data by upsampling technique and using data coded with the same value in first content analysis. All performance index was more than 70%, which showed that psychological distance can be measured well.

Keywords: SNS Marketing, Psychological Distance, Natural Language Processing, Machine Learning, Support Vector Machine

* Research Fellow, Department of Biomedical Informatics, School of Medicine, Ajou University.
** Corresponding Author, Student, Department of Computer Science, College of Engineering, Dankook University

◐ 저 자 소 개 ◑

이 성 원 (seongwon.lee.16@gmail.com)

아주대학교 의료정보학과 연구강사로 재직 중이다. 연세대학교 경영대학원에서 정보시스템 전공 석-박사를 취득하였다. 연구 관심 분야는 Health IT, 빅데이터 분석, 디지털 정보 행태 등이며, Journal of Associated for Information Systems, Asia-Pacific Journal of Information Systems, Information Systems Review 등의 저널에 논문을 발표하였다.

김 진 혁 (withhhhh@gmail.com)

단국대학교 컴퓨터과학과 4학년에 재학 중이며, 관심분야는 모바일 및 IoT 기기의 취약점 분석과 IT인프라의 데이터 수집과 분석이다.

논문접수일 : 2018년 11월 05일 **게재확정일 : 2018년 12월 26일**
1차 수정일 : 2018년 12월 24일