http://dx.doi.org/10.14400/JDC.2016.14.3.165

SNS에서의 언어 간 감성 차이 연구: 6개 언어를 중심으로

김형호*, 장필식**
세한대학교 정보물류학과*, 세한대학교 정보물류학과**

Differences in Sentiment on SNS: Comparison among Six Languages

Hyung-Ho Kim*, Phil-Sik Jang**
Dept. of Information & Logistics of Sehan University*
Dept. of Information & Logistics of Sehan University**

요 약 본 연구의 목적은 SNS 활용에 있어 사용자 언어 간 감성의 평균차이가 있는지를 검증하는 것이다. 가장 많이 이용되는 SNS 중 하나인 트위터를 대상으로, 영어, 독일어, 러시아어, 스페인어, 터키어 및 네덜란드어 등 6개 언어로 작성된 약 2억 개 트윗을 스트리밍 API를 이용하여 수집하였으며, SentiStrength를 이용하여 주관적/객관적 비율, 감성강도, 긍정/부정 비율, 리트윗 횟수 및 경계불투과도 등에 대한 분석을 시행하고, 트위터를 통한 감성표현의 경향성과 변동을 파악하였다. 분석결과, 언어권에 따라 주관적/객관적 트윗 비율과 긍정/부정 트윗 비율이 각각 통계적으로 유의한 차이가 있는 것으로 나타났다($p<0.001$). 또한, 언어의 종류는 감성강도와 경계 불투과도 그리고 리트윗 횟수에 통계적으로 유의한($p<0.001$) 영향을 미치는 것으로 파악되었다. 이러한 결과는 SNS를 활용한 감성분석에 있어 언어, 문화 별 경향성 및 수준차이를 반드시 고려하여야 한다는 것을 보여준다.

주제어 : SNS, 감성분석, 트위터, 언어·문화차이, 감성강도

Abstract The purpose of this study was to explore the differences in sentiment on social networking sites among six languages (English, German, Russian, Spanish, Turkish and Dutch). A total of 204 million tweets were collected using Streaming API. Subjective/objective ratio, sentiment strength, positive/negative ratio, number of retweets and boundary impermeability were analyzed with SentiStrength to estimate the trends of emotional expression via Twitter. The results showed that subjective/objective ratio and the positive/negative ratio of tweets were significantly different by languages ($p<0.001$). And, there were significant effects of language on sentiment strength, boundary impermeability and the number of retweets ($p<0.001$). The results also indicate that the cross-cultural, language differences should be taken into account in sentiment analysis on SNS.

Key Words : SNS, Sentiment Analysis, Twitter, Cultural Difference, Sentiment Strength

* 이 논문은 2016년도 세한대학교 교내 연구비 지원에 의하여 씌여진 것임
Received 1 February 2016, Revised 26 January 2016
Accepted 20 March 2016, Published 28 March 2016
Corresponding Author: Phil-Sik Jang(Sehan University)
Email: phil@sehan.ac.kr

ISSN: 1738-1916

1. 서론

최근 SNS는 개인 간의 단순한 의사소통 수단을 넘어, 사회, 경제, 문화, 정치 등 다양한 분야로 활용의 범위가 급격히 확대되고 있으며, SNS 데이터를 분석, 활용하고자 하는 다양한 노력들이 경주되고 있다[1,2]. 이러한 노력들 중 최근 주목받고 있는 기술 중 하나가 감성분석(sentiment analysis)이다[3]. 텍스트 마이닝(text mining) 기법을 통해 이루어지는 감성분석은 특정 언어에 종속될 수밖에 없어서, 각 언어 별로 정확도를 높이기 위한 다양한 알고리즘개발이 이루어지고 있다[4]. 하지만 이러한 기술적 부분과는 별개로, SNS 메시지가 전 세계에 걸쳐 실시간으로 확산된다는 점에서 감성분석의 적용 시, SNS사용자의 언어・문화적 특성과 경향성에 대한 고려가 필요할 것으로 생각된다.

Boyd & Ellison[5]은 사용자의 문화적 특성과 환경이 컴퓨터를 매개로한 의사소통(CMC: Computer Mediated Communication)에 영향을 미치며, SNS 또한 의사소통과 상호작용 활동에 있어 사용자의 문화적 특성에 영향을 받는다고 주장하였다. 또한 SNS에서의 자기표현(self- presentation)은 미국-한국[6], 미국-중국[7], 미국-일본 학생 간[8]에 유의한 차이가 있다는 연구결과도 발표되었다. 이와 같이 SNS 활용에서의 문화적 차이에 대한 연구가 최근 등장하고 있지만, 이들 연구들은 실험실 내 소수 특정 집단(주로 대학생)의 수동적 피실험자들을 대상으로 두 개 문화 간의 자기표현 차이만을 연구 대상으로 하고 있다. 수집된 트위터 메시지를 대상으로 아프리카, 인디아, 영국, 미국 등 네 개 지역을 비교한 연구[9]도 있으나, 영어로 작성된 메시지만을 대상으로 하였으며, 하루 동안의 감정(emotion)변화만 분석하여, 다양한 언어 문화권 특성은 다루지 못한 한계점을 가진다. 즉, 특정 이슈나 인물, 제품에 대한 만족도, 유통 및 물류서비스 등에 대하여 다양한 문화, 언어를 포함한 감성 비교분석이 이루지기 위해서는 각 문화, 언어 별로 긍정/부정 표현에 있어 기본적 경향성이 있는지, 있다면 그 변동성이 어느 정도인지에 대한 분석이 선행되어야 하는데, 이와 관련된 실증적 연구는 아직 찾아보기 어렵다.

본 연구에서는 대표적 SNS인 트위터(Twitter)를 대상으로, 2014년 11월부터 2015년 4월까지 영어, 스페인어 등 6개 언어로 작성된 204,569,362개 메시지(트윗)를 수집하였으며, 각 언어별로 감성분석을 시행하고, 트위터를 통한 감성표현의 경향성과 변동을 파악하였다.

2. 연구 대상 및 방법

본 연구에서는 트위터 메시지의 언어권별 감성특성을 파악하기 위해 다음과 같이 감성분석에 직・간접적으로 연관되는 변수들에 대한 분석을 시행하였다.

주관적 메시지 비율 : 전체 메시지를 주관적(subjective), 객관적(objective) 메시지로 구분할 때 주관적 메시지의 백분율

감성강도(Sentiment Strength) : 각 메시지 내 포함된 감성의 강도. -4~4점 scale 척도로 가장 강한 긍정적 감성은 4점, 가장 강한 부정적 감성은 -4점으로 분류

긍정비율 : 주관적 메시지 중 긍정메시지 백분율

리트윗(Retweet) 수 : 트위터의 정보 확산 정도를 가늠하는 척도[10]로써, 메시지(트윗) 당 리트윗 횟수

경계 불투과도(Boundary Impermeability) : 트위터 이용자의 자기 노출정보가 얼마나 많은 사람들에게 노출가능한가의 척도[11]. 감성에 영향을 받는 것으로 알려져 있다.

$$Boundary\,Impermeability = \frac{N_{followers}}{N_{followers} + N_{following}}$$

트윗 수집에는 Twitter Inc에서 제공[12]하는 스트리밍 API(Streaming Application Programming Interface)를 이용하였으며, 스트리밍 데이터는 매 15분마다 분류, 처리하고 MongoDB 3.2(MongoDB Inc.)[13] 데이터베이스에 저장하였다. 감성분석에는 빠른 분석 속도와 비교적 높은 정확도로 최근 여러 연구[14]에 활용되고 있는 SentiStrength[15]를 이용하였다. SentiStrength는 다양한 언어에의 적용이 가능한데, 본 연구에서는 현재까지 검증된 6개 언어(영어, 스페인어, 독일어, 러시아어, 네덜란드어, 터키어)를 대상으로 하였다.

3. 분석 결과

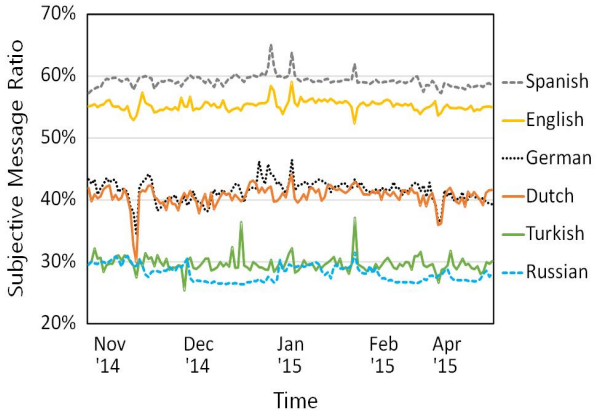
3.1 언어권 별 주관적(subjective) 트윗 비율

각 언어권 별로 주관적 트윗 비율에 차이가 있는지를 검증하기 위해 χ^2 검정을 시행하였으며 검정 결과, 유의수준 0.001에서 언어권에 따라 주관적/객관적 트윗 비율에 통계적으로 유의한 차이가 있는 것으로 나타났다.

<Table 1> Pearson's Chi-squared Test (Subjective Message vs Language)

Language	N(subjective)	N(total)	Ratio(subjective)
Dutch	614,988	1,518,077	40.5%
English	75,660,434	136,735,207	55.3%
German	758,973	1,764,000	43.0%
Russian	4,515,763	15,685,078	28.8%
Spanish	23,154,096	39,262,220	59.0%
Turkish	2,785,189	9,604,781	29.0%
χ^2 = 6,914,600 df = 5 p-value < 2.2e-16			

영어와 스페인어 사용자의 경우 주관적 감정, 감성을 표현한 트윗은 각각 평균 55.3%, 59.0%로써 객관적 내용보다 주관적 내용의 트윗이 좀 더 많음을 보여주고 있다. 데이터 수집 기간 동안 일별 평균추이를 언어권별로 나타낸 결과는 [Fig. 1]과 같은데, 각 언어권별로 큰 변동 없이 일정 수준을 유지하는 것으로 보인다.



[Fig. 1] Daily Means of Subjective Message Ratio

3.2 언어권 별 감성강도(sentiment strength)

트윗의 감성 강도(-4~4)를 종속변수로 하고 언어권을 독립변수로 하는 분산분석을 실시하였으며(Table 3), 유의수준 0.001에서 언어군에 따라 감성강도 평균이 통계적으로 유의한 차이를 보이는 것으로 나타났다.

<Table 2> Sentiment Strength according to Languages

Language	Mean	SD	SE.	99% CI
Dutch	0.06	0.98	8.0.e-04	2.1.e-03
English	0.21	1.22	1.0.e-04	2.7.e-04
German	0.28	0.99	7.4.e-04	1.9.e-03
Russian	0.16	0.68	1.7.e-04	4.4.e-04
Spanish	0.44	1.37	2.2.e-04	5.6.e-04
Turkish	0.12	0.78	2.5.e-04	6.5e-04

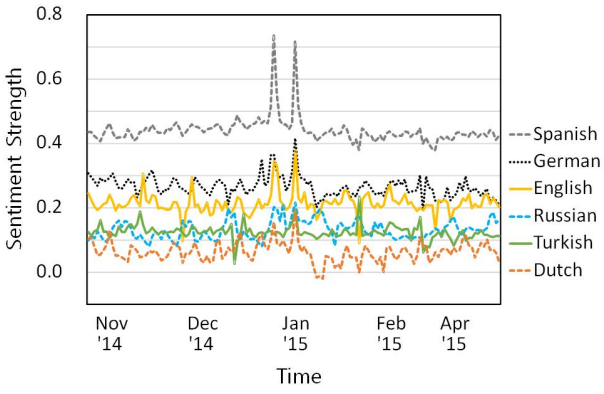
<Table 2>에서 볼 수 있는 것처럼, 스페인어로 작성된 트윗의 감성강도 평균이 가장 높은(가장 긍정적인) 것으로 나타났으며, 네덜란드어와 터키어로 작성된 트윗의 강도가 가장 낮은 것으로 관찰되었다.

수집 기간 동안 일별 평균된 감성강도의 언어군 별 추이를 살펴보면[Fig. 2], 12월 25일이나 1월 1일과 같은 특정한 날짜에 큰 변동이 있는 것으로 보이나, 전체적으로 언어군 별로 일정한 수준을 유지하는 것으로 관찰된다.

<Table 3> ANOVA Summary Table : Sentiment Strength by Languages

Source	df	SS	MS	F	Pr(>|t|)
L	5	1,848,321	369,664	259,299	<2.2e-16***
Residuals	204,569,357	291,639,743	1		
Total	204,569,362	293,488,064			

L=Language *p<.05, **p<.01, ***p<.001



[Fig. 2] Daily Means of Sentiment Strength

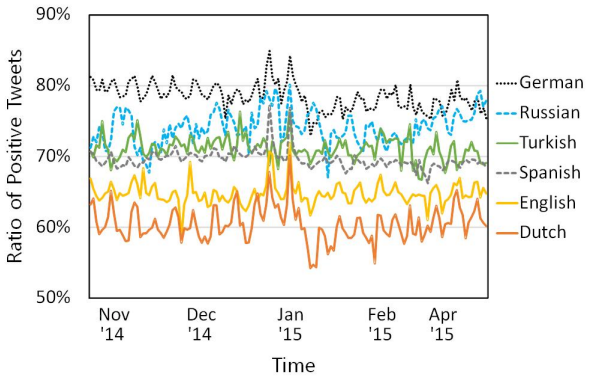
3.3 주관적 트윗의 긍정 비율

주관적인 내용인 것으로 분류된 트윗들 중 긍정적 트윗의 비율이 언어권에 따라 다른지를 판별하기 위해 시행한 피어슨 χ^2 검정 결과는 <Table 4>와 같다.

<Table 4> Pearson's Chi-squared Test
(Positive tweets ratio vs Language)

Language	N(positive)	N(subjective)	Ratio(positive)
Dutch	371,591	614,988	60.4%
English	49,050,795	75,660,434	64.8%
German	595,361	758,973	78.4%
Russian	3,462,373	4,515,763	76.7%
Spanish	16,122,972	23,154,096	69.6%
Turkish	1,982,555	2,785,189	71.2%
χ^2 = 492,970 df = 5 p-value < 2.2e-16			

검정 결과, 언어권에 따라 긍정적 트윗의 비율이 유의수준 0.001에서 통계적으로 유의한 차이를 보이는 것으로 나타났다. 독일어의 경우, 주관적 트윗 중 긍정의 비율이 평균 78.4%로 가장 높은 것으로 나타났으며, 네덜란드어는 평균 60.4%로 가장 낮게 나타났다. 일별 평균된 긍정비율의 언어군 별 추이는 [Fig. 3]와 같은데, 일간 변동은 있으나 언어 군 별로 비율차이는 어느 정도 일정하게 유지되는 것으로 보인다.



[Fig. 3] Daily Mean Ratio of Positive Tweets

3.4 감성 극성과 경계 불투과도

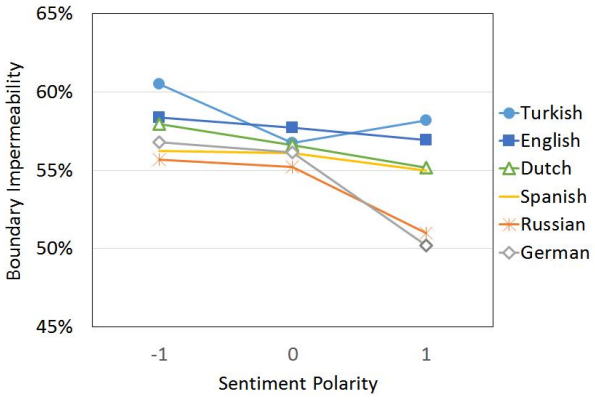
트위터 메시지의 감성극성(polarity)과 언어군이 경계불투과도에 영향을 미치는 지를 판별하기 위해, 경계불투과도를 종속변수로 하고 언어군과 감성극성을 독립변수로 실시한 ANOVA 분석 결과는 <Table 5>와 같다.

분석결과 언어군과 감성극성에 따라 경계 불투과도 평균은 유의수준 0.001에서 통계적으로 유의한 것으로 나타났으며, 두 독립변수의 교호작용 또한 유의한 것으로 나타났다. [Fig. 4]에서 볼 수 있는 것처럼, 언어군 별에 차이는 있으나 전체적으로 감성극성이 부정(-1)인 경우보다 감성극성이 긍정(+1)일 때 경계불투과도가 낮은 것으로 판단된다. 이 결과는 Walton and Rice[11]의 연구결과와 일치하는 것이며, 이러한 연구결과의 활용에 있어 사용자 언어에 따른 문화적 차이 또한 고려해야할 요인임을 보여주고 있다.

<Table 5> ANOVA Summary Table : Boundary Impermeability by Language, Sentiment Polarity

Source	df	SS	MS	F	Pr(>|t|)
L	5	23,727	4,745	85,533	<2.2e-16***
SP	2	7,411	3,705	66,786	<2.2e-16***
L×SP	10	5,304	530	9,559	<2.2e-16***
Residuals	203,406,913	11,284,931	0.1		
Total	203,406,930	11,321,373			

L=Language
SP=Sentiment Polarity
*p<.05, **p<.01, ***p<.001



[Fig. 4] Boundary Impermeability vs Language, Sentiment Polarity

3.5 감성 강도와 리트윗

리트윗 수가 트위터 메시지의 감성강도와 언어군에 따라 영향을 받는지를 살펴보기 위해 리트윗 수를 종속변수로 하고 감성강도와 언어군을 독립변수로 하는 분산분석을 실시한 결과는 <Table 6>과 같다. 분석결과, 유의수준 0.001에서 감성강도와 언어군, 그리고 두 변수의 교호작용이 통계적으로 유의한 영향을 미치는 것으로 나타났다. [Fig 5]에서 볼 수 있는 것처럼, 언어에 따라 감성강도와 리트윗 수 간의 관계가 다른 양상을 보일 수 있는 것으로 나타났는데, 예를 들어 영어로 작성된 트윗의 경우 감성강도가 높아질수록(긍정적일수록) 리트윗 횟수가 많아지는 경향이 관찰되지만, 네덜란드어와 러시아어

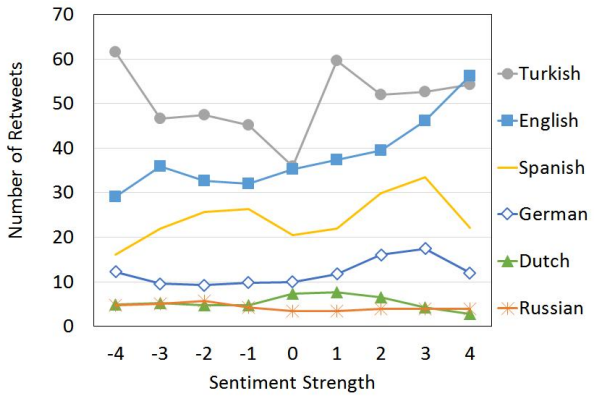
에서는 이러한 경향이 나타나지 않았다.

<Table 6> ANOVA Summary Table : N of Retweets by Language, Sentiment Strength

Source	df	SS	MS	F	Pr(>|t|)
L	5	2.0e+10	4.0e+09	6,302.37	<2.2e-16***
NR	8	1.4e+09	1.7e+08	272.761	<2.2e-16***
L×NR	40	1.1e+09	2.8e+07	44.474	<2.2e-16***
Residuals	204,569,309	1.3e+14	6.3e+05		
Total	204,569,362	1.3e+14			

L=Language
NR=N of Retweets

*p<.05, **p<.01, ***p<.001



[Fig. 5] Number of Retweets vs Language, Sentiment Strength

4. 결론

본 연구에서는 각 언어 별로 긍정/부정 표현에 있어 기본적 경향성이 있는지, 있다면 그 변동과 추이가 어떠한지를 파악하기 위해, 6개 언어로 작성된 204,569,362개 트윗(tweet)에 대한 감성분석을 실시하였으며, 결과를 요약하면 다음과 같다.

첫째, 언어권에 따라 주관적/객관적 트윗비율이 통계적으로 유의한($p<0.001$)차이를 보였다. 영어와 스페인어 사용자들의 트윗이 터키와 러시아어에 비해 주관적 트윗 비율이 높은 것으로 나타났다.

둘째, 언어에 따라 감성강도 평균이 통계적으로 유의한($p<0.001$) 차이를 보였는데, 스페인어로 작성된 트윗의 감성강도가 터키어와 네덜란드어로 작성된 트윗에 비해 높게(긍정인 것으로) 나타났다.

셋째, 주관적 트윗 중 긍정 트윗의 비율이 언어권에 따라 통계적으로 유의한($p<0.001$)차이를 보이는 것으로 나타났다. 비율이 가장 높은 독일어와 가장 낮은 네덜란드어의 비율차이는 18% 정도였다.

넷째, 위에서 언급된 트윗의 주관적/객관적 비율, 감성강도 평균 및 긍정트윗 비율은 일별 변동은 있으나 언어별로 일정한 수준을 지속적으로 유지 하는 것으로 판단된다.

다섯째, 경계불투과도는 감성극성과 언어군에 통계적으로 유의한($p<0.001$) 영향을 받는데, 긍정적일수록 경계불투과도는 낮아지는 경향이 있는 것으로 관찰되었다.

여섯째, 리트윗 횟수는 감성강도와 언어에 따라 유의한($p<0.001$) 영향을 받는 것으로 나타났으며, 교호작용도 유의한 것으로 관찰되었다.

이러한 결과는 SNS를 활용한 특정 이슈나 인물, 제품 만족도, 유통 및 물류서비스 등에 대한 감성분석에 있어 언어문화 별 경향성 및 수준차이를 반드시 고려하여야 한다는 것을 보여준다.

본 연구에서는 SNS 중 트위터에 한정하여 6개 언어·문화권에 대한 반년 미만의 시간적 추이를 분석하였다. 본 연구결과를 바탕으로, 페이스북을 포함한 다양한 SNS와 블로그를 대상으로 좀 더 다양한 언어군에 대한 추가 연구가 이루어진다면, 다양한 언어, 문화권을 아우르는 활용성 높은 감성분석 체계를 구축하는데 도움이 될 것으로 기대된다.

ACKNOWLEDGMENTS

This Paper was supported by the Sehan University Research Fund in 2016.

REFERENCES

[1] K. Choi, J. A. Yoo, “A reviews on the social network analysis using R”, Journal of the Korea Convergence Society, Vol. 6, No. 1, pp. 77-83, 2015.

[2] J. Y. Go, K. H. Lee, “SNS disclosure of personal information in M2M environment threats and countermeasures”, Journal of the Korea Convergence Society, Vol. 5, No. 1, pp. 29-34, 2014.

[3] B. Pang, and L. Lee, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, Vol. 1. No. 2, pp. 1-135, 2008.
[4] P. S. Jang, "Study on principal sentiment analysis of social data", Journal of The Korea Society of Computer and Information, Vol. 19, No. 12, pp.49-56, 2014.
[5] D. M. Boyd and N. B. Ellison, N. B. "Social network sites: Definition, history, and scholarship", Journal of Computer-Mediated Communication, Vol. 13, No. 1. pp. 210-230, 2007
[6] J. J. Yoo, D. Kim and J. Moon, "Exploring cross-cultural differences in self-presentation and self-disclosure in social Networking sites: A comparison of korean and american SNS users", Journal of Advertising and Promotion Research, Vol. 1, No. 2, pp. 77-118, 2012.
[7] S. Luo, "Cross-cultural differences between American and Chinese college students on self-disclosure on social media. Graduate Theses and Dissertations", p.1-70, Iowa State University. 2014.
[8] K. Omori and M. Allen, "Cultural differences between american and japanese self-presentation on SNSs", International Journal of Interactive Communication Systems and Technologies (IJICST), Vol. 4, Issue 1. DOI: 10.4018/ijicst. 2014010104, 2014.
[9] S. A. Golder and M. W. Macy, "Diurnal and seasonal mood vary with work, sleep and daylength across diverse cultures." Science, 333, pp. 1878 - 1881, 2011.
[10] J. Y. Lee, P. S. Jang, "Effects of message polarity and type on word of mouth through SNS (Social Network Service)", The Journal of Digital Policy & Management, Vol. 11, No. 6, pp. 129-135, 2013.
[11] S. C. Walton and R. E. Rice, "Mediated disclosure on Twitter: The roles of gender and identity in boundary impermeability, valence, disclosure, and stage", Computers in Human Behavior, Vol. 29, pp. 1465-1474, 2013.
[12] "The Streaming APIs", © 2016 Twitter, Inc., https://dev.twitter.com/streaming/overview (Jan 5, 2016)
[13] "MongoDB 3.2", MongoDB, Inc., https://www.mongodb.org/ (Jan 5, 2016)
[14] M. Thelwall, K. Buckley. & G. Paltoglou, "Sentiment strength detection for the social Web", Journal of the American Society for Information Science and Technology, Vol. 63, No. 1, pp. 163-173, 2012.
[15] "SentiStrength", http://sentistrength.wlv.ac.uk/, (Jan 25, 2016)

김 형 호(Kim, Hyung Ho)

・1989년 2월 : 경희대학교 전자계산공학과 (공학사)
・1992년 8월 : 경희대학교 전자계산공학과 (공학석사)
・1998년 3월 : 일본 게이오대학 계산기과학과 (공학박사)
・1998년 3월 ~ 현재 : 세한대학교 정보물류학과 교수
・관심분야 : 신경회로망, 감성분석, 물류정보시스템, 해운물류
・E-Mail : hhkim@sehan.ac.kr

장 필 식(Jang, Phil Sik)

・1990년 2월 : 서울대학교 조선공학과(공학사)
・1992년 2월 : KAIST 산업공학과 (공학석사)
・1998년 8월 : KAIST 산업공학과 (공학박사)
・1997년 9월 ~ 현재 : 세한대학교 정보물류학과 교수
・관심분야 : HCI, 물류정보시스템, 감성분석
・E-Mail : phil@sehan.ac.kr