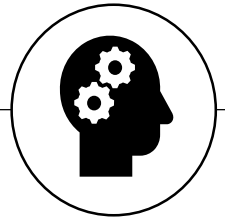
소셜미디어 데이터 마이닝을 활용한 스마트 스피커 제품에 대한 VOC 분석

박영재 건국대 산업공학과
dudwo1128@gmail.com



Index

연구 목적

스마트 스피커

- 구글 홈, 아마존 에코 등으로 대표되는 스마트 스피커는 무선스피커에 인공지능 비서를 접목
- 이용자와의 상호작용을 통하여 정보 알림, IOT (Internet of things) 기기 조작 등 다양한 기능을 제공하는 제품
- 2014년 아마존 에코를 시작으로 본격적으로 스마트 스피커 시장이 형성되었고 그에 따른 제품 및 서비스가 증가하고 있음
- 인공지능, 음성 인식 기술들의 발달로, 스마트 스피커를 위시한 스마트 홈, IOT 제품 및 서비스 개발이 활발히 이루어지고 있음

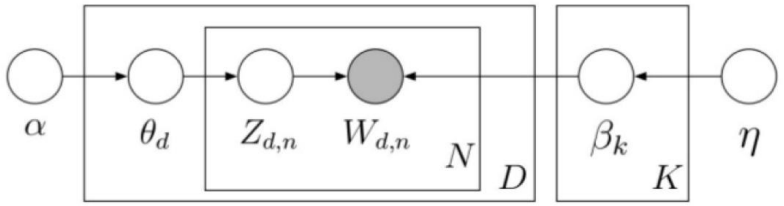
관련 연구

- 스마트 홈은 거주지의 모니터링과 자동화를 가능하게 함 (Cook 2012)
- 스마트 홈 장애인과 노약자와 같이 특별한 도움이 필요한 이용자들을 대상으로 차별화된 시장을 형성 할 수 있을 것으로 보여짐 (Smirek, Zimmermann, and Beigl 2016).
- 기업들은 IOT에 대한 잠재성을 높게 평가하여 IOT에 대한 적극적인 투자를 신중하게 고려하고 있음 (Lee and Lee 2015)
- 스마트 스피커 시장규모는 2015년 3억 6천만 달러 (약 4천 300억 원)에서 2020년 20억 달러 (약 2조 3천억원)로 6배 이상 커질 것으로 예측 (김영대 2017)
- 스마트 스피커와 같은 이머징 제품의 경우에는 기술주도형 (Technology-driven)으로 개발되는 경향을 지니며 (Haines et al. 2007)

관련 연구

- 시장에서의 소셜 미디어 영향력은 지대하여 소셜 미디어를 통한 커뮤니케이션은 마케팅에서 중요한 이슈로서 활용됨 (Mangold and Faulds 2009)
- 제품 사용자들의 정보 교환의 창구로서의 역할을 한다 (Wang, Yu, and Wei 2012)
- 소셜 미디어는 대용량인 동시에 실시간 정보이므로 시장에 대한 기업의 즉각적인 대응을 가능하게 함 (Sakaki, Okazaki, and Matsuo 2010)
- 소셜 미디어는 고객 입장에서 제품 개발 방향 또는 비즈니스 인사이트를 얻을 수 있으며, 이러한 관점을 이용하여 소셜 미디어를 대상으로 다양한 연구들이 진행되고 있다. 소셜 미디어를 CRM에 적용할 수 있음 (Malthouse et al. 2013)
- 소셜 미디어 데이터를 분석하여 박스 오피스 산업에서의 비즈니스 인텔리전스에 활용함 (Lu, Wang, and Maciejewski 2014)

이론적 배경



K – total number of topics
β_k – topic, a distribution over the vocabulary
D – total number of documents
Θ_d – per-document topic proportions
N – total number of words in a document (it fact, it should be N_d)
$Z_{d,n}$ – per-word topic assignment
$W_{d,n}$ – observed word
α, η – Dirichlet parameters

- Several **inference algorithms** are available (e.g. sampling based)
- A few **extensions** to LDA were created:
  - Bigram Topic Model

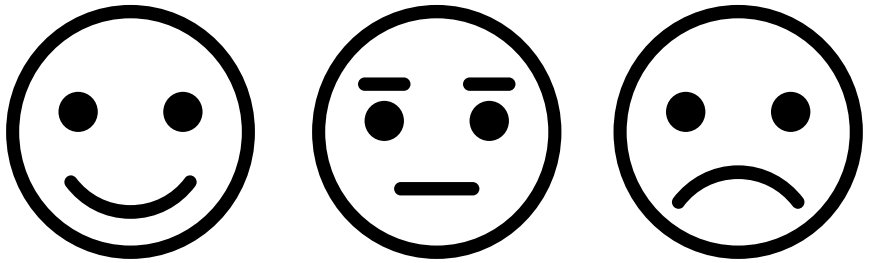
토픽 모델링

- 기계 학습과 자연어 처리 분야에서 문서 집합의 주제를 파악하기 위해 사용하는 방법
- 문서에 자주 등장하는 단어들의 통계를 이용한 분석 모델
- 문서가 단어로 구성되어 있다는 점을 이용하여 비슷한 주제의 문서에는 비슷한 키워드가 등장한다는 전제로 이루어지는 주제 클러스터링 방법

LDA(Latent Dirichlet Allocation)

- 토픽 모델링 중에서 가장 많이 사용
- Overfitting 문제가 없으며 새로운 문서에 대해서도 쉽게 일반화가 가능함

이론적 배경



Positive Neutral Negative

AlchemyAPI™

감성 분석

- 자연어 처리(Natural Language Process), 텍스트 분석 등의 활용
- 설문조사, 후기, 소셜 미디어 데이터 등의 Voice of Customer에 적용
- Lexicon-based, Rule-based, Deep learning-based의 3가지 방법이 존재함

IBM Watson : Alchemy API

- Java, Python 등의 다양한 프로그래밍 언어 지원
- 9개의 언어의 텍스트 이해
- 키워드에 대한 긍정/ 부정을 판단
- -1과 1사이의 점수를 부과

연구 절차

본 연구에서는 소셜 미디어 데이터 마이닝을 활용하여 제품 계획을 하고자 하며,
이를 달성하기 위해 Jeong, Yoon, and Lee (2017)의 방법을 활용함

1. 소셜 미디어 기반의 소비자 리뷰 데이터를 수집하였으며, 해당 데이터는 미국 온라인 커뮤니티 레딧 (Reddit)에서 아마존 에코 데이터 약 11만 건과 구글 홈 데이터 약 7만 건을 사용하였음

2. 토픽모델링을 통하여 고객들에 의해 자주 언급되는 제품 토픽을 정의하고 토픽의 중요도를 결정함

3. 감성 분석을 통하여 제품 토픽의 만족도를 산출함

4. 기회분석 알고리즘을 이용하여 고객의 관점에서 제품 토픽의 중요도와 만족도를 평가함

5. 경쟁 제품의 제품 토픽 비교 분석을 실시하며, 이를 통해 시장의 의견을 반영한 제품의 개선 기회 및 개발 방향을 제시함

연구 절차

소비자 리뷰 데이터 수집

구글 홈, 에코 Reddit 데이터 약 18만 건 수집

제품 토픽 결정

키워드 추출 & 토픽 모델링

제품 토픽의 중요도 결정

개별 문서가 토픽에 속하는 비율

제품 토픽의 만족도 결정

IBM Watson을 통한 감성 분석

경쟁 제품의 토픽 간의 기회 비교, 분석

제품 개발 기회 도출

연구 절차

소비자 리뷰 데이터 수집

created_utc	subreddit_id	parent_id	body
1501761254	t5_3enp4	t1_dl3sj2a	Six, and she went on to list them.
1502453120	t5_3enp4	t1_dlgy67c	I have had the shortcuts for approx two weeks now but not the services
1503587971	t5_3enp4	t1_dm2cvyj	I don't have that listed. Is there a way to install it?
1503750205	t5_3enp4	t1_dm4e69f	This is a great feature.
1504197743	t5_3enp4	t1_dmdmas9	You want to know the time from MY watch? TOO BAD.

Reddit

- 소셜 뉴스 웹사이트
- 다양한 주제에 대한 피드 업데이트
- 해당 주제에 관련한 서브 레딧이 존재함
- 본 연구에서는
  'Googlehome'과 'Echo' 서브 레딧 데이터 활용

데이터 수집

- 구글 BigQuery를 활용하여 데이터 수집
- 작성자, 댓글 데이터 수집
- Echo 데이터 116,464건,
  Googlehome 데이터 77,740건 수집

연구 절차

제품 토픽 결정

키워드 추출, 정제 및 감성 분석

- IBM Watson Alchemy API 활용
- 키워드, 해당 문서에서 언급 횟수, 긍정/부정, 점수 추출
- Document Frequency, Term Frequency, 불용어 고려하여 키워드 정제
- Echo, GoogleHome과 관련되지 않은 키워드 제거
- Github https://github.com/dudwo1128/Python_Project/blob/master/Echo_KWD_Extract.py

Keywords	numbers	Sentiment	Score
integration	1	positive	0.69917
thing	1	negative	-0.47906
iHeart Radio playlists	1	positive	0.560155
color band	1	positive	0.80525
wrong music	1	negative	-0.42256
voice training	1	positive	0.332312
mics	1	positive	0.829191
ton	1	positive	0.560155
instances	1	negative	-0.42256
direction	1	positive	0.80525
couple	1	negative	-0.42256
time	1	positive	0.560155

연구 절차

제품 토픽 결정

토픽모델링

- Echo 키워드 2051개,
  GoogleHome 키워드 1416개 활용
- 토픽의 수는 토픽 사이의 코사인 유사도를 통하여 결정
- Echo 토픽 23개, GoogleHome 토픽 23개 활용
- Echo, GoogleHome 관련 조사를 통한
  토픽 Labeling 실시
- GitHub
  https://github.com/dudwo1128/Python_Project/blob/master/TopicNumber_Decision.py

Topic	1st Keywo	2nd Keywo
Voice Control	spotify	alarm
Echo Connect	phone	wall
Light Control	light	ifttt
Echo Plus	song	setup
Echo Dot	echo dot	alexa app
Audio/Video Remote Control	account	remote
Notification	tunein	notificatior
Smart Plug	timer	command
Smart Home Automation	house	living roon
Alexa	alexa	harmony h
Alexa Skill	order	receiver
Communications	bluetooth	data
Home Control	alexa	wifi
Question & Answer	question	comment

연구 절차

제품 토픽의 중요도 결정

- 각 토픽은 웹 데이터를 이용하여 정의되어 소비자들이 해당 제품에 대해 직접적으로 관심있는 제품
- 토픽에 해당하는 문서가 많다는 것은 해당 토픽에 기대하는 수준이 높음을 의미
- 동시에 해당 토픽의 중요도가 높은 것으로 해석
- 개별 문서가 토픽에 속할 확률의 합을 토픽의 중요도로 해석
- 각 토픽의 중요도를 10점 척도로 정규화

	D_1	D_2	D_3	D_4	⋯	D_n
$Topic_1$	0.010	0.051	0.001	0.154		
$Topic_2$	0.112	0.508	0.004	0.103		
$Topic_3$	0.235	0.002	0.014	0.144		⋮
$Topic_4$	0	0.004	0.475	0		
⋮				D		
$Topic_n$			⋯			⋱

Topics-Documents matrix

Summation

	Stock
$Topic_1$	146.7
$Topic_2$	87.9
$Topic_3$	122.3
$Topic_4$	93.2
⋮	⋮
$Topic_n$	103.6

Contribution stock for topics

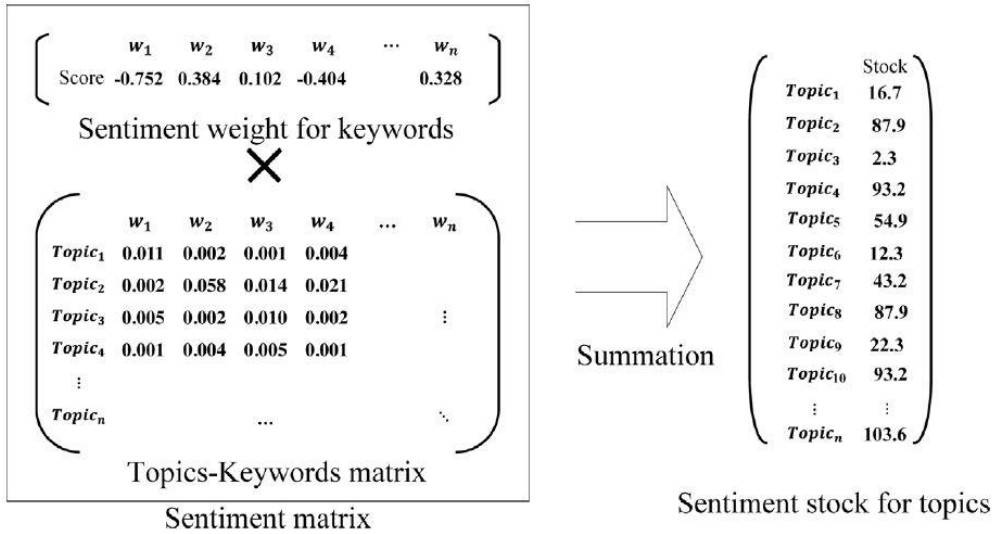
$$CS_t = \sum_{i=0}^{\#ofDocuments} TDMatrix_{t,i}, \text{ Where t= Topic\#}$$

$$Importance_i = 10 \times \frac{CS_i - CS_{Min}}{CS_{Max} - CS_{Min}}$$

연구 절차

제품 토픽의 만족도 결정

- 토픽 모델링의 결과인 토픽과 문서 사이의 확률분포와, 문서와 키워드 사이의 감성점수 활용
- 확률분포와 감성 점수를 벡터 곱하여 감성 점수분포를 가짐
- 감성점수분포는 이론적으로 -1과 1사이의 값을 가짐
- 감성점수를 10점 척도로 변환하여 만족도 결정



$$SS_t = \sum_{i=0}^{\#ofDocuments} SentimentMatrix_{t,i}, \text{ Where t= Topic\#}$$

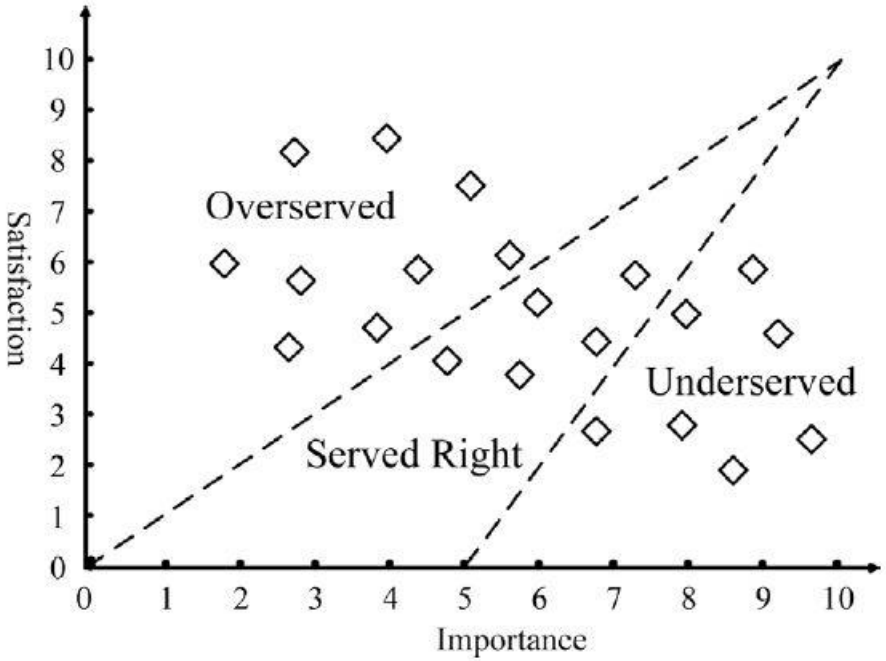
$$Satisfaction_i = 10 \times \frac{SS_i - SS_{Min}}{SS_{Max} - SS_{Min}}$$

연구 절차

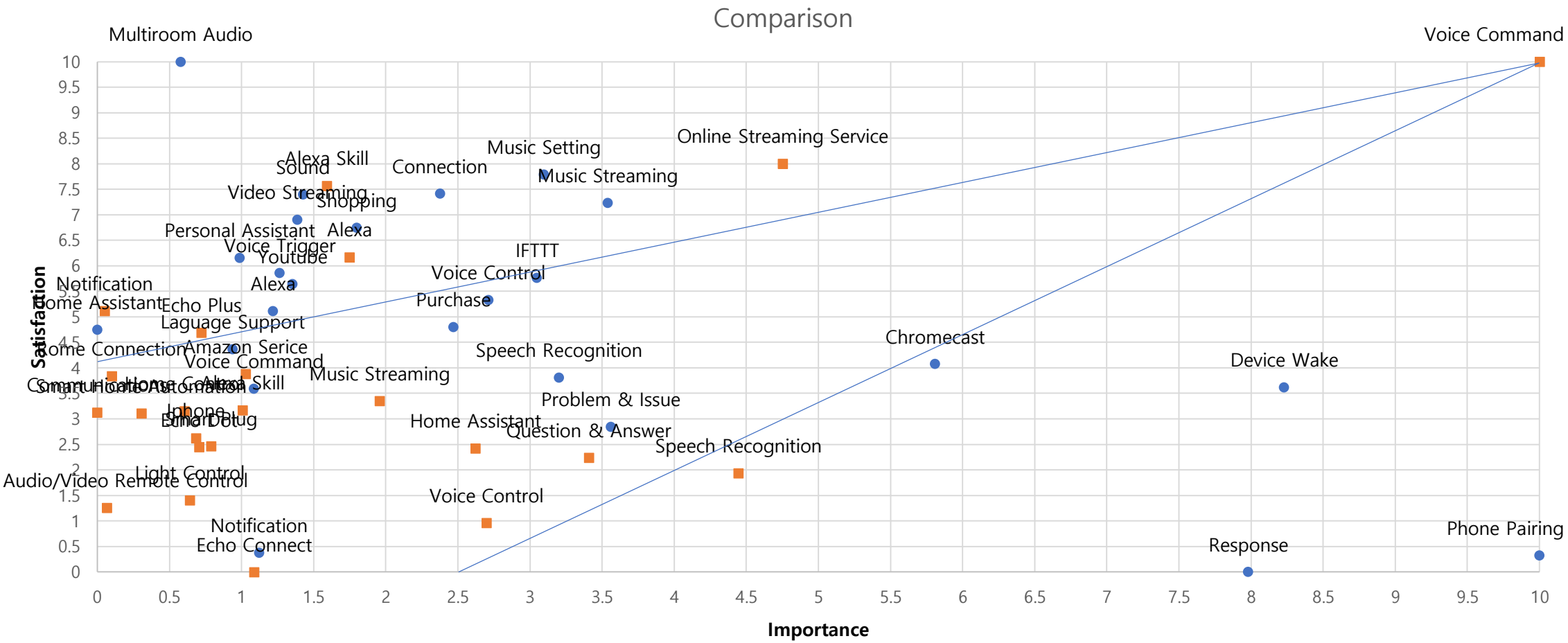
경쟁 제품의 토픽 간의 기회 비교, 분석

Opportunity Algorithm

- 10점 척도로 계량화 된 개별 토픽의 중요도와 만족도를 2차원에 맵핑
- Opportunity Algorithm을 이용하여 각 토픽에 대해 개선의 기회 파악
- 중요도와 만족도의 값에 따라서 3가지 영역으로 구분할 수 있음
- Under-Served 영역에 해당하는 토픽들을 개선의 기회가 있는 토픽으로 판단함



결론


Comparison
Multiroom Audio
Voice Command
Online Streaming Service
Music Setting
Alexa Skill
Sound
Connection
Music Streaming
Video Streaming
Shopping
Personal Assistant
Alexa
Voice Trigger
Youtube
IFTTT
Voice Control
Alexa
Notification
Purchase
Echo Plus
Laguage Support
Chromecast
Amazon Serice
Speech Recognition
Device Wake
Voice Command
Music Streaming
Problem & Issue
Home Assistant
Question & Answer
Speech Recognition
Light Control
Audio/Video Remote Control
Voice Control
Notification
Echo Connect
Phone Pairing
Response
Satisfaction
Importance

참고 문헌

Cook, Diane J. 2012. 'How smart is your home?', *Science*, 335: 1579-81.
Haines, Victoria, Val Mitchell, Catherine Cooper, and Martin Maguire. 2007. 'Probing user values in the home environment within a technology driven Smart Home project', *Personal and Ubiquitous Computing*, 11: 349-59.
Jeong, Byeongki, Janghyeok Yoon, and Jae-Min Lee. 2017. 'Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis', *International Journal of Information Management*.
Lee, In, and Kyoochun Lee. 2015. 'The Internet of Things (IoT): Applications, investments, and challenges for enterprises', *Business Horizons*, 58: 431-40.
Lu, Yafeng, Feng Wang, and Ross Maciejewski. 2014. 'Business intelligence from social media: A study from the vast box office challenge', *IEEE computer graphics and applications*, 34: 58-69.
Malthouse, Edward C, Michael Haenlein, Bernd Skiera, Egbert Wege, and Michael Zhang. 2013. 'Managing customer relationships in the social media era: Introducing the social CRM house', *Journal of interactive marketing*, 27: 270-80.
Mangold, W Glynn, and David J Faulds. 2009. 'Social media: The new hybrid element of the promotion mix', *Business Horizons*, 52: 357-65.
Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. 2010. "Earthquake shakes Twitter users: real-time event detection by social sensors." In *Proceedings of the 19th international conference on World wide web*, 851-60. ACM.
Smirek, Lukas, Gottfried Zimmermann, and Michael Beigl. 2016. 'Just a smart home or your smart home–a framework for personalized user interfaces based on eclipse smart home and universal remote console', *Procedia Computer Science*, 98: 107-16.
Wang, Xia, Chunling Yu, and Yujie Wei. 2012. 'Social media peer communication and impacts on purchase intentions: A consumer socialization framework', *Journal of interactive marketing*, 26: 198-208.