Social media mining for product planning:

A product opportunity mining approach based on topic modeling and sentiment analysis

비즈니스인포매틱스학과
석사 2기 전소영

목 차

Introduction

1 Introduction

기업은 고객에게 신상품 혹은 개선된 제품을 제공하기 위하여 고객의 목소리에 주의를 기울여야 함.

연구개발 및 마케팅에 대한 접근 방식은 고객의 니즈 분석에 초점을 두었는데,
특히 고객 분석을 통해 잠재적인 제품의 필요성을 일찍 발견하는 것이 제품 개발 및 개선 프로세스에 가장 중요.
새로운 고객의 니즈를 일찍 파악하면 경쟁 업체가 쉽게 모방할 수 없는 고객 맞춤 브랜드 형성에 유리.

그러나, 오늘날 단축된 제품 수명과 글로벌 비즈니스 환경으로 인해 최근 고객의 요구 사항은 더욱 역동적이고 복잡해짐.

1 Introduction

이전보다 복잡해진 고객의 니즈를 발견하고자
많은 기업들이 블로그, Twitter, Facebook, Reddit과 같은 소셜 미디어를 적극적으로 활용.

소셜 미디어를 통해 고객의 개인적인 의견과 제품과 관련된 최신 데이터들 제공 받을 수 있다는 것이 장점.

실제로 미국 성인의 86%와 유럽 성인의 79%가 소셜 미디어를 사용하므로
소비자 의사결정 프로세스 및 마케팅 커뮤니케이션을 위해 소셜 미디어 데이터는 분석 가치를 지님.

Social media mining approach for Product opportunities

O Opportunities algorithm + T Topic modeling + S Sentiment analysis

- T 소셜 미디어 내 소비자 리뷰 데이터를 토픽 모델링하여 제품 토픽을 인지 후, 수집된 데이터를 기반으로 제품 토픽의 **중요도**를 측정함.
- S 감성 분석을 통해 각 제품 토픽의 **만족도**를 계산함.
- O 기회 알고리즘을 적용하여 각 제품 토픽의 **기회 점수**를 측정 후, 제품 토픽에 대한 평가 진행.

Contributions

I. 제안된 접근 방식은 소셜 미디어 데이터를 사용하여 제품 키워드에 대해 잠재력을 평가하므로 고객 중심의 제품 개발 방향의 우선 순위를 지정하는 데 도움을 줄 수 있음.

II. Opportunity analysis은 제품 뿐만 아니라 서비스 및 제품-서비스 시스템에도 적용 가능. 이는 본 연구가 제안하는 접근 방식이 도메인 중립적이며 소셜 미디어에서 수집 된 텍스트 데이터에만 의존하기 때문.

III. 급변하는 고객 요구에 대응하는 실시간 고객 모니터링 도구로서의 상품 기획 분석 시스템 개발의 기반이 될 수 있음.

Theoretical background

Opportunity algorithm

- ✓ 본 연구에서는 기회 알고리즘을 통해 각 제품 키워드가 고객 중심 관점에서 개선될 잠재력을 식별함.
- ✓ Ulwick[1)]이 제안한 기회 알고리즘은 충족되지 않은 요구의 우선 순위를 지정하는데 사용되는 방법으로, 고객 관점의 중요도와 만족도를 기반으로 0부터 10까지로 측정됨.
- ✓ 해당 키워드가 중요하지만, 충분히 만족되지 않을 때 혁신의 기회가 존재한다는 가정을 바탕으로 실행되기 때문에, 가장 중요하지만 가장 적게 충족되는 키워드일수록 가장 높은 우선 순위를 받음.

$$\text{Opportunity} = \text{Importance} + \text{Max}(\text{Importance} - \text{Satisfaction}, 0)$$

1) Ulwick, A. W. (2005). What customers want: Using outcome-driven innovation to create breakthrough products and services, Vol. 71408673. New York: McGraw-Hill.

Proposed methodology

3 Proposed methodology

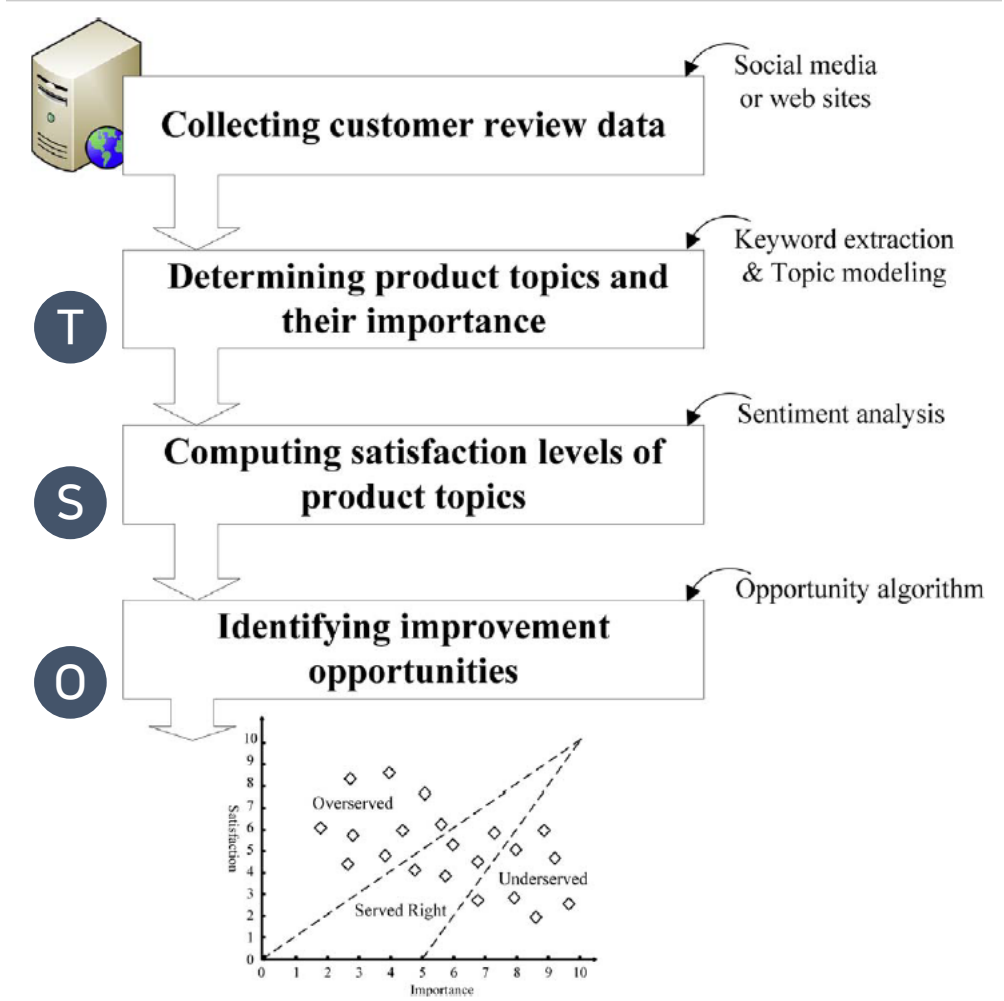


Fig. 2. Overview of the proposed approach.

3 Proposed methodology

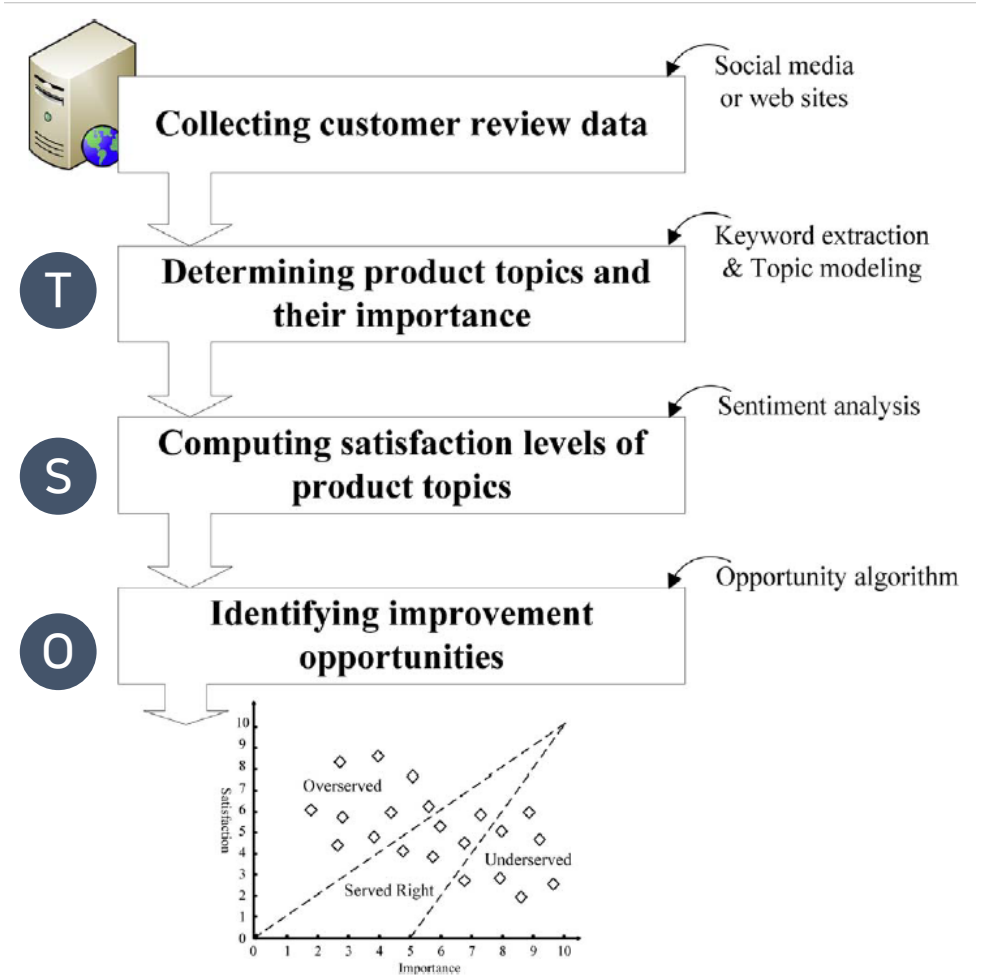


Fig. 2. Overview of the proposed approach.

Step 1 Data gathering and preprocessing

- ✓ 개방형 어플리케이션 프로그래밍 인터페이스 (API)를 제공하는 Twitter, Reddit과 같은 소셜 미디어 선택.
- ✓ 소셜 미디어 내 특정 제품에 대한 리뷰 데이터 수집.
- ✓ 고객의 온라인 리뷰에서 키워드를 추출하여 각 리뷰 데이터셋을 구성함. 키워드 목록을 구성하는 과정에서 이모티콘, 의성어, 관련 없는 단어는 목록에서 제외.

Step 2 Identifying product topics and their importance

$$\text{Cosinesimilarity(A, B)} = \cos(\theta) = \frac{A \cdot B}{AB} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

- ✓ LDA 기반의 토픽 모델링 기법을 활용하여 제품 토픽 추출.
- ✓ 이 과정에서 코퍼스(말뭉치)는 온라인 리뷰, 단어 사전은 키워드 목록을 의미. 그리고, 적절한 토픽 수의 경우. 토픽 모델링에 의해 생성된 모든 토픽-단어 분포 벡터 pair 간 평균 코사인 유사도를 활용하여 최적 토픽 수를 측정 (elbow method).

- ✓ 앞서 추출한 제품 토픽들을 바탕으로 사용자가 **각 제품 토픽을 언급하는 정도**를 계산하여 제품 주제의 중요도를 측정.
- ✓ 모든 고객 리뷰에 대한 **각 제품 토픽의 분포 확률의 합**은 수집된 말뭉치(=온라인 리뷰)에서 해당 제품 토픽의 중요성을 의미함.
- ✓ 그리고 0-10의 척도로 정규화 진행함.

	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

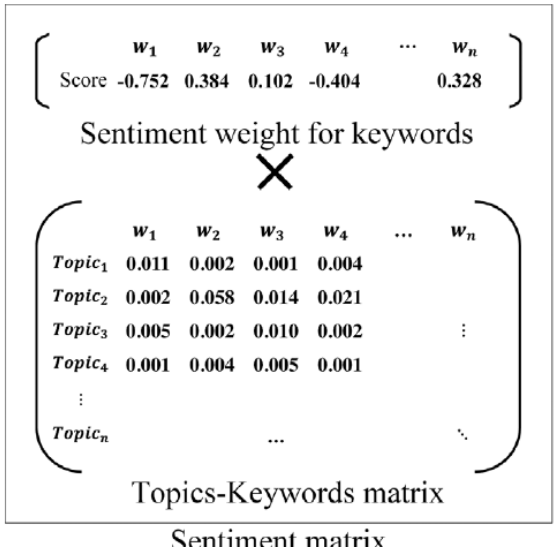
$$CS_t = \sum_{i=0}^{\#ofDocuments} TDMatrix_{t,i}, \text{ Where t= Topic\#}$$

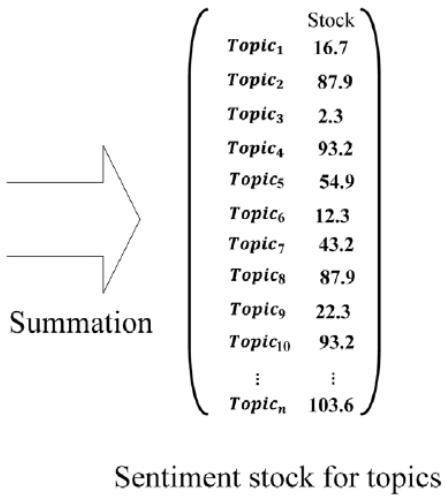
$$Importance_i = 10 \times \frac{CS_i - CS_{Min}}{CS_{Max} - CS_{Min}}$$

Step 3 Computing the satisfaction level of product topics

- ✓ **감성분석**을 활용하여 각 키워드의 평균 감성 점수를 정의하고 키워드별 평균 감성점수로 구성된 배열 생성.
- ✓ 이전 단계에서 추출된 토픽-키워드 매트릭스에 키워드 감성 벡터 곱하여 감성 가중치-토픽 키워드 매트릭스 구성.

- ✓ 해당 제품의 토픽을 구성하는 키워드의 가성치를 합산하여 제품 토픽의 만족도를 산출함.
- ✓ 중요도와 마찬가지로, 만족도도 1-10로 변환.





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

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

Step 4 Identifying product opportunities using the opportunity algorithm

- ✓ 기회 알고리즘에 따르면, 중요도는 높고 만족도는 낮을수록 높은 기회 점수를 얻음.
- ✓ 기회 landscape map을 구성하여 3부분으로 나눔 (served-right, over-served, and underserved).

- ✓ Served-right: 중요도에 비해 적절한 만족도를 가지는 키워드
- ✓ Over-served: 중요도에 비해 높은 만족도를 가지는 키워드
- ✓ Underserved: 중요도에 비해 낮은 만족도를 가지는 키워드 → 제품 개선에 활용 가능한 토픽!

- ✓ 높은 기회 점수를 가진 토픽들 중에서 특히, 감성 가중치가 높은 키워드들을 중심으로 만족도를 높여 제품을 개선할 수 있음

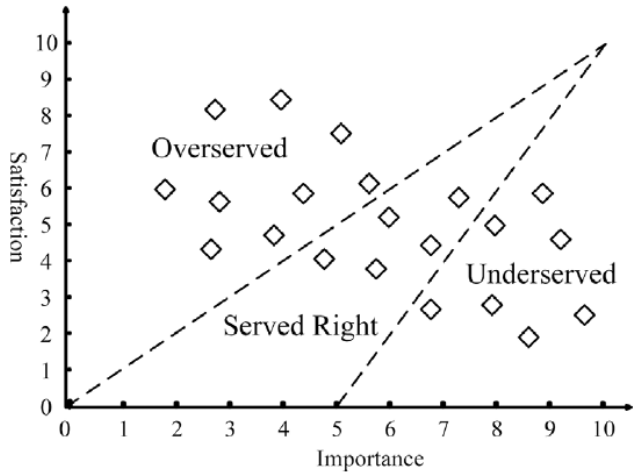
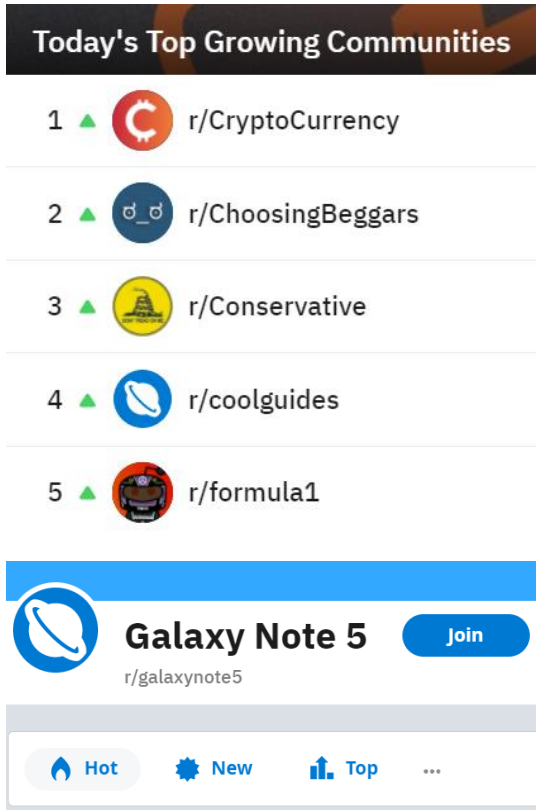


Fig. 5. Schematic of opportunity landscape maps.

Case study: Samsung galaxy note 5

4 Case study: Samsung galaxy note 5

Step 1 Data gathering and preprocessing



reddit

- ✓ 여러 소셜 미디어 중에서 **Reddit**을 선택함. Twitter의 경우 140 문자 수 제한이 있으므로 주로 # (해시태그)를 통해 의견을 전달한다는 단점이 있으나, Reddit은 주제별로 subreddit을 따로 갖고 있다는 점에서, 특정 제품에 대한 리뷰를 수집하기 용이함.
- ✓ **삼성 갤럭시 노트5의 subreddit** (https://www.reddit.com/r/galaxy_note5)에서 2014년 11월 7일부터 2016년 1월 31일까지 작성된 23,614개의 documents 데이터 수집 (2255 posts, 21,539 comments). + SGN5와 무관한 광고, 노이즈 게시물 제거.
- ✓ **Rapid Automatic Keyword Extraction 비지도 알고리즘**을 통해 키워드 추출함. 세 단계에 걸친 노이즈 제거를 통해 최종 3539개의 키워드 선정. 해당 키워드가 등장하지 않는 document는 제거함. 그래서 최종 11,123 documents가 분석에 사용됨.

Step 2 Identifying product topics

- ✓ Netminer를 활용하여 keyword frequency matrix를 구성하고 LDA 진행함.
- ✓ 토픽 pair 간 평균 **코사인 유사도**를 활용하여 적절한 토픽 개수를 정함. 본 연구에서는 65개일 때 가장 낮은 유사도였으므로 65개로 정함. LDA로 추출된 65개의 토픽들을 Topic1~Topic65로 지칭함.
- ✓ 토픽별로 구성하는 키워드들의 분포와 관련 리뷰들 확인.

Table 1
Part of keyword-frequency matrix of SGN5 .

Document ID	4G	4GB	Exynos	OS	RAM	64GB	battery	BT	car
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	0	0	0	0
5	1	1	0	2	0	1	1	1	1
6	0	0	0	0	0	0	0	1	0
7	1	1	1	5	2	0	1	0	3
8	0	0	0	9	0	0	6	0	3
10	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	1	0	0
13	0	0	0	0	0	0	1	0	0
15	0	0	0	0	0	0	0	0	0

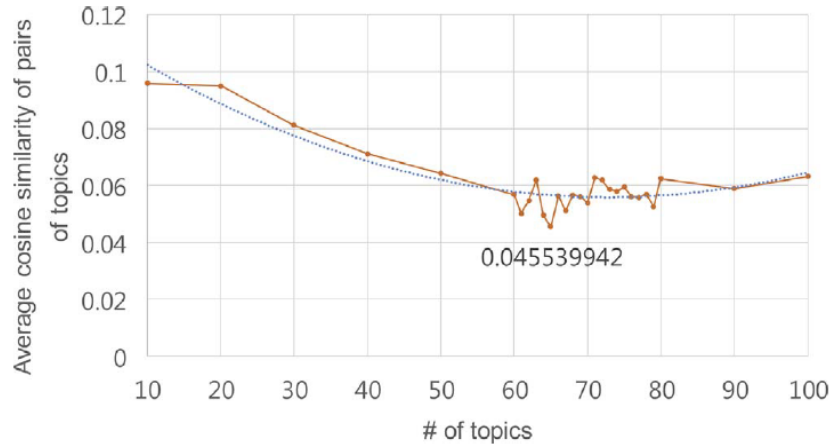


Fig. 7. Average cosine similarities of pairs of SGN5 topics.

Table 4
Explanation of some topics .

Topics	Major keywords (Topic contribution); Representative data
Design	design (0.118), material (0.029), Material Design (0.014), Marshmallow (0.013), OS (0.009), stock Android (0.008), UI (0.008) "Great implementation of Material Design. Clean and simple. That's all I need a texting app for.", "That too! With that crazy glass on glass design, I'd put a case on it right away. Better to save the design rather than see it shatter in a drop. I get that."
Calling	LTE (0.159), calling (0.079), VoLTE (0.037), WiFi-Calling (0.036), battery (0.049), advanced calling (0.013), video calling (0.005) "Keep in mind if you don't get the T-Mobile version you will NOT get VoLTE WiFi calling or band 12 voice (since its VoLTE.)", "I turned off volte. I'm keeping WiFi calling on though. I can do without volte but not WiFi calling."
Software update	update (0.344), software (0.053), software update (0.025), security (0.012), updating (0.006), Marshmallow update (0.005), security update (0.004) "There was a new update that redid the app. Have you updated? My widget crashed after the update and I had to put up a new widget based on the updated version.", "New software update Just downloading a new software update for the Note 5 on AT & T. Anyone know what it's for? This is different from the one a few weeks ago."
Samsung pay	pay (0.331), SamsungPay (0.210), Samsung Pay (0.209), android pay (0.029), NFC (0.022), pay app (0.012), payments (0.011) "Thanks! It actually turned out that I had to first turn on NFC, switch to the Samsung Pay app from Android Pay, which then allowed me to access the menu for Samsung Pay, to then go to the settings and turn off Simple Pay. Otherwise known as a bug.", "I know, but the terminal says it accepts Apple Pay, so then wouldn't it also be able to accept Android Pay, Google Wallet, or the NFC part of Samsung Pay? (Samsung pay is both NFC and MST)"
Camera	camera (0.314), picture (0.156), camera app (0.015), shutter (0.009), front-facing camera (0.005), photograph (0.005), lens (0.005) "How to turn off tap to take pictures? Is it possible to turn off tap to take pictures on the front facing camera mode?", "With the stock camera app I can only change video resolution. I would imagine a third party camera app could do what you wanted."
Battery life	battery (0.261), battery-life (0.234), better battery (0.012), better battery life (0.009), poor battery life (0.004), good battery life (0.004), great battery life (0.003), "Note 5 Bad Battery Life I am getting very bad battery life on the Verizon Note 5. Android System is using most of my battery and then it is cell standby. I am currently at 53% with only 1 h of on screen time. What can I do to fix this?", "Wow. I would never in a million years strip my phone so much just to gain a few more minutes of battery life. You don't need to do any of this to have great battery life."

Step 3 Computing importance and satisfaction degree of product topics

- ✓ LDA 과정에서 keyword frequency matrix로부터 1)Document-topics matrix, 2)Topics-Keywords matrix를 추출하는데, 이 두 매트릭스를 통해 중요도와 만족도 측정함.
- ✓ Document-topics matrix을 통해 각 문서 내 토픽 비중을 합하여 **토픽의** Contribution stock 계산.

~ 본 연구에서 가장 높은 중요도는 Samsung pay, 낮은 중요도는 Accessory 토픽이었음.

- ✓ AlchemyAPI로 **딥러닝 기반의 감성분석** 진행하였는데, 3539개의 키워드에 대한 평균 감성 점수를 측정. 평균 감성 점수는 감성 가중치로 고려하고 앞서 구한 Topics-Keywords와 곱함. ~ 만족도

Keywords	Weight	Keywords	Weight	Keywords	Weight
display	0.0158	charging	−0.0511	camera	0.1450
Exynos	−0.0083	wireless charging	−0.1010	design	0.2438
battery	−0.2334	5.7-inch display	0	s-pen out	0.3816
removable battery	−0.2453	64-bit architecture	0.3645	battery size	−0.0856
SD card	−0.2963	AMOLED	0.5689	finger print scanner	−0.5141
upgrade	0.2027	auto-focus	0.0334	charger	−0.1597
accessories	−0.0810	Battery life	−0.0457	OS systems	0

Topic	Importance	Satisfaction
Samsung Pay	10.0000	7.8185
Fast charge	7.0599	5.4720
Detect pen2	4.3805	0.1384
Pen out	4.6130	0.6877
Battery life	4.2233	1.7818
Expandability	3.6873	2.1963
Software update	4.6186	4.9924
Detect pen1	2.2850	0.0000
Charger	4.3514	4.4286
Screen glass	3.9796	4.8418

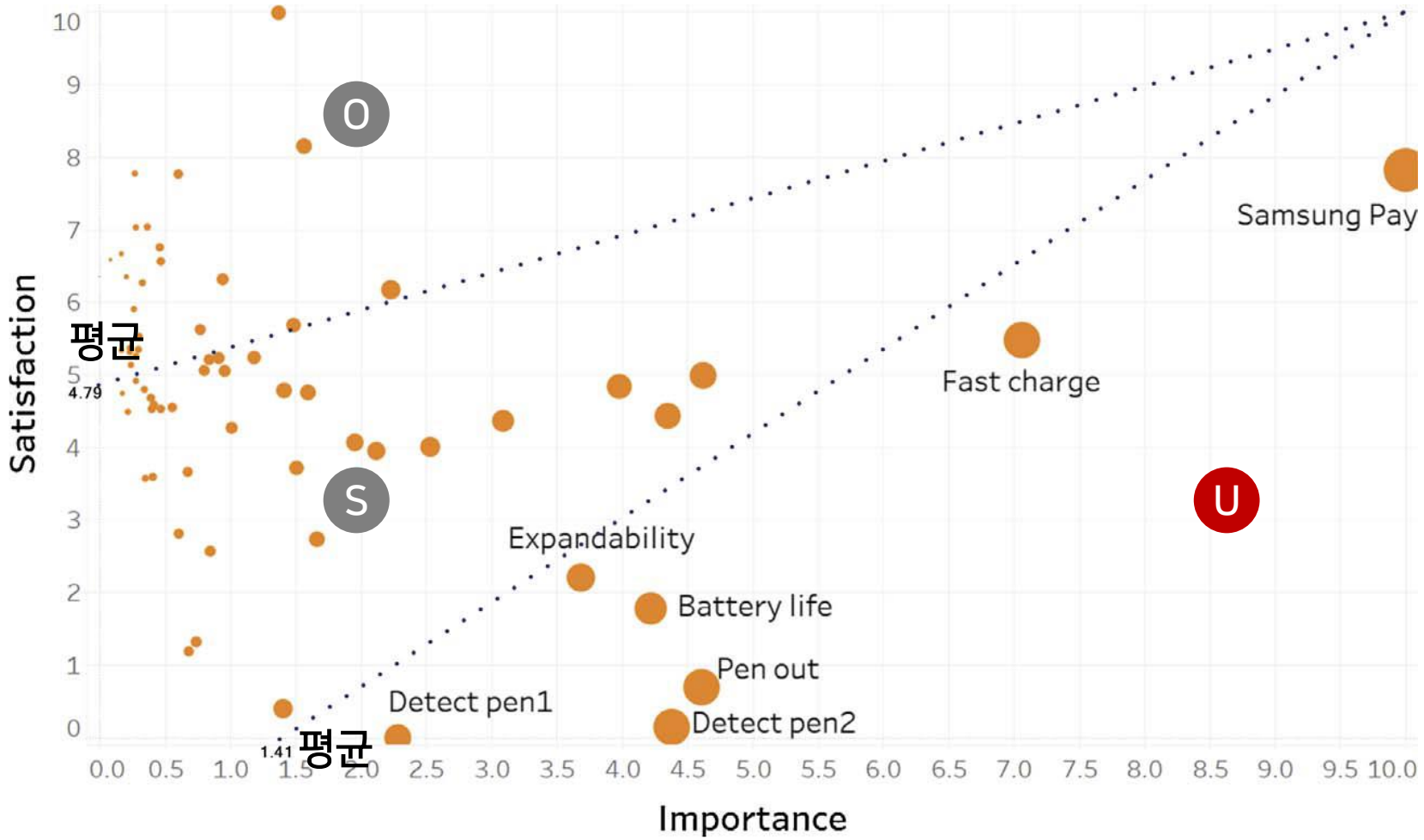
Step 4 Identifying product opportunities

✓ 앞서 측정한 중요도와 만족도를 활용하여 SGN5의 **기회 점수**를 구하고 **Landscape map**을 그릴 수 있음.

Table 7
Opportunity score of product topics.

Topic	Opportunity	Topic	Opportunity	Topic	Opportunity
Samsung Pay	12.1816	Widgets	1.1853	Hardware Spec.	0.3668
Fast charge	8.6478	Fingerprint	1.0090	Calling	0.3514
Detect pen2	8.6225	Custom Rom	0.9580	Optimization	0.3409
Pen out	8.5384	Internal storage	0.9452	Edge display	0.3276
Battery life	6.6648	Icon & wallpaper	0.9126	Multi-tasking	0.3065
Expandability	5.1783	Physical buttons	0.8501	Default application	0.2974
Software update	4.6186	Device connect	0.8403	Data Transfer	0.2818
Detect pen1	4.5700	Video record	0.7990	Case	0.2805
Charge cable	4.3514	Theme	0.7744	Game	0.2768
Screen glass	3.9796	Network dropped	0.7387	New OS feature	0.2706
Screenshot	3.0895	SMS	0.6854	Stylus	0.2666
Lock screen	2.5322	Battery usage	0.6723	SIM	0.2406
Write on screen	2.4151	Battery drain	0.6077	Play store	0.2351
Wireless charge	2.2300	OS upgrade	0.6013	Design	0.2306
Wi-Fi	2.1173	Location	0.5573	Screen off memo	0.2182
Screen(AMOLED)	1.9574	OTA	0.4712	Samsung Pay on ATM	0.2052
Emoji	1.6663	Warranty & Repair	0.4699	Sound	0.1751
Music App	1.5975	Galaxy note5	0.4645	Material	0.1687
Camera	1.5629	Screen resolution	0.4159	Google Play	0.1680
Multi windows	1.5081	Accessory	0.4057	Hardware performance	0.0822
Hand write	1.4836	UX	0.4033	Accessory	0
Touch wiz	1.3710	E-mail	0.3966	–	–
–	–	–	–	**Average**	**1.7098**

- ✓ Served-right: 중요도에 비해 적절한 만족도를 가지는 키워드
- ✓ Over-served: 중요도에 비해 높은 만족도를 가지는 키워드
- ✓ Underserved: 중요도에 비해 낮은 만족도를 가지는 키워드 → 제품 개선에 활용 가능한 토픽!



✓ Landscape map 내 Underserved에 해당하는 토픽들에 기여하는 **키워드들에 대한 감성 점수** 확인
✓ 이중에서 **긍정 감성 점수가 높은 키워드는 강화하고, 부정 감성 점수가 낮은 키워드는 보완할 것!**

Samsung Pay		Fast charge		Detect pen2	
Keyword	Sentiment	Keyword	Sentiment	Keyword	Sentiment
NFC payment	−0.00032814	charger	−0.03531872	S-Pen sensor	−0.00334226
NFC terminals	−0.00020921	wireless Charger	−0.00471534	pen detection	−0.00112448
non NFC terminal	−0.00010881	Samsung Wireless Charger	−0.00089303	S-Pen Backwards	−0.00044851
card read error	−0.00005386	Wireless Charging Paused	−0.00024550	Broken S-pen Sensor	−0.00031429
error messages	−0.00005105	Charging Paused message	−0.00006287	spring mechanism	−0.00010550
Samsung Pay Rebate	0.00005936	awesome features	0.00004939	screen-off memo	0.00005606
Loop pay	0.00008111	wireless quick charging	0.00008577	s-pen out.	0.00007288
new samsung pay	0.00011817	Samsung Wireless Charging	0.00009715	screen-off	0.00015206
Samsung pay promotion	0.00014520	quick charging	0.00009911	S-Pen menu	0.00022321
Samsung Pay	0.01376659	fast wireless charger	0.00297911	S-Pen work	0.00249590
Battery life		**Expandability**		**Software update**	
Keyword	**Sentiment**	**Keyword**	**Sentiment**	**Keyword**	**Sentiment**
battery-life	−0.13852779	SD card	−0.02764752	factory reset	−0.00029786
poor battery life	−0.00249269	MicroSD	−0.01246854	Manually update	−0.00018607
Package Disable	−0.00027939	removable battery	−0.01191136	TouchWiz skin	−0.00009915
background service	−0.00022641	IR blaster	−0.00366371	error messages	−0.00008012
unnecessary process	−0.00018517	LGG4	−0.00050291	overheating	−0.00006232
battery optimization	0.00016404	selfies	0.00009094	Samsung support	0.00005581
Amoled	0.00017123	OTG	0.00010466	camera software	0.00006450
Screen-on time	0.00025804	battery-pack	0.00012827	optimization	0.00011683
awesome battery life	0.00029115	portable power banks	0.00015185	new Samsung pay	0.00016881
iPhone5	0.00122801	cloud storage	0.00015369	software update	0.00323952

Conclusions

5 Conclusions

본 연구를 통해 소셜 미디어 내 소비자 리뷰 데이터로 제품 개발 기회를 정량화하고 제품 개발 지침을 제시할 수 있다.

분석 활용 방안

- 해당 연구는 특정 제품에 대한 리뷰들을 수집하여 직접적인 제품 개발 기회를 모색했다는 점에서 차별성을 가지지만, SGN5와 같은 글로벌 제품에 대해서만 하위 레딧들이 존재한다는 점이 한계….

- 해당 연구의 경우 특정 제품에 대한 데이터를 분석했다는 점에서 진행될 분석과 차이가 있음. 그러나, 해당 연구가 제안한 **기회 알고리즘+토픽 모델링+감성분석** 접근법을 그대로 적용해볼 수 있음. 특정 제품이 아니라 육아, 패션과 같은 소비자들의 일상적인 카테고리에 대하여 분석이 가능하다고 생각함.

- 그러나, 연구에서 진행될 tool을 그대로 사용하기 보다는 Python 코드로 분석을 진행할 예정.

- RAKE 알고리즘을 통한 별도의 키워드 추출 과정을 거칠 수도 있지만, 세부 토픽 모델링과 군집화를 거쳐 Topic와 Keywords로 구분하는 방법 또한 해당 접근법에 적용할 수 있다고 생각.

Thank you ☺