International Journal of Information Management 48 (2019) 280–290

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis

Byeongki Jeong[a], Janghyeok Yoon[a,*], Jae-Min Lee[b]

[a] Department of Industrial Engineering, Konkuk University, Seoul, Republic of Korea
[b] Future Information Research Center, Korea Institute of Science and Technology Information, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Product opportunity
New product development
Social media mining
Opportunity algorithm
Topic modeling
Sentiment analysis

ABSTRACT

Social media data have recently attracted considerable attention as an emerging voice of the customer as it has rapidly become a channel for exchanging and storing customer-generated, large-scale, and unregulated voices about products. Although product planning studies using social media data have used systematic methods for product planning, their methods have limitations, such as the difficulty of identifying latent product features due to the use of only term-level analysis and insufficient consideration of opportunity potential analysis of the identified features. Therefore, an opportunity mining approach is proposed in this study to identify product opportunities based on topic modeling and sentiment analysis of social media data. For a multifunctional product, this approach can identify latent product topics discussed by product customers in social media using topic modeling, thereby quantifying the importance of each product topic. Next, the satisfaction level of each product topic is evaluated using sentiment analysis. Finally, the opportunity value and improvement direction of each product topic from a customer-centered view are identified by an opportunity algorithm based on product topics' importance and satisfaction. We expect that our approach for product planning will contribute to the systematic identification of product opportunities from large-scale customer-generated social media data and will be used as a real-time monitoring tool for changing customer needs analysis in rapidly evolving product environments.

1. Introduction

Commercial firms need to pay particularly close attention to customer voices to provide customers with new or improved products. Accordingly, approaches to research and development (R & D) and marketing have naturally placed considerable emphasis on customer needs analysis. Using these approaches, attempts have been made to co-create value with customers, because firms are directly or indirectly engaged with customers who want to satisfy their needs by purchasing or hiring various solutions such as products and services (Griffin & Hauser, 1993;Silverstein, Samuel, & DeCarlo, 2013). In particular, early identification of new or attractive product opportunities from customer analysis is considered to be the first and critical requirement in the product development or product improvement process (Van Kleef, van Trijp, & Luning, 2005). This is because early identification of such opportunities enables a firm to create a unique and customized relationship, which cannot be easily copied by competing firms and accordingly leads to the competitive positioning of the firm in the value chain (Park & Yoon, 2015).

However, today's reduced product lifecycles and globalized business environments have made recent customer needs more dynamic and complex than ever before. In this regard, social media such as blogs, Twitter, Facebook, Reddit, and other social networking services provide a good source of plentiful and real-time data related to customer opinions (Brooks, 2015). As a tool for social interaction and information exchange, social media is quickly becoming a channel for exchanging customer-generated open voices about products and the number of such voices stored in social media has increased explosively (Wang,Yu, & Wei,2012). In fact, it was found that 86% of adults in the United States and 79% of adults in Europe use social media (Sverdlov, 2012). Therefore, social media data are considered worthy of analysis for the purpose of supporting consumer decision-making processes and marketing communications (Kietzmann, Hermkens, McCarthy, & Silvestre, 2011).

Park & Yoon, 2015
Geum et al., 2015
Geum, Lee, Lee, & Park, 2015
Kang & Park, 2014

In response to this issue, a social media mining approach for product opportunities is proposed in the present study using an opportunity algorithm based on topic modeling and sentiment analysis. In this

* Corresponding author.
E-mail address: janghyoon@konkuk.ac.kr (J. Yoon).

http://dx.doi.org/10.1016/j.ijinfomgt.2017.09.009
Received 13 April 2017; Received in revised form 11 August 2017; Accepted 26 September 2017
Available online 12 October 2017
0268-4012/

approach, product topics (i.e. the product topics that are currently discussed by product consumers) are identified by topic modeling of online customer reviews in social media. The importance level of each product topic is then calculated by the topic modeling analysis based on the product topic's knowledge stock. In addition, the satisfaction level of each product topic is computed using sentiment analysis. Finally, the opportunity score of each product topic is evaluated by applying the opportunity algorithm based on the importance and satisfaction of the product topic. This enables analysts to direct further product development based on the keywords of the topic with high opportunity potential. To show the operation of this approach, it was applied to social media data in which the Samsung Galaxy Note 5 (SGN5) was reviewed.

The contributions of the present study are threefold. First, the proposed approach is able to evaluate the potential opportunity of product topics for improvement using social media data. This feature could help prioritize product development directions for customer-centered product planning. Second, our opportunity analysis approach has the potential for application to not only products but also services and product-service systems. This is because our approach is domain-neutral and depends only on the textual data collected from social media. Third, this approach contributes to the systemizing of product opportunity analysis processes and therefore can act as a real-time customer monitoring tool to cope with rapidly changing customer needs while becoming a basis for the development of an intelligence system that assists product planners.

The organization of this study is as follows. We present the groundwork of this study, followed by our approach and its application to identify product improvement opportunities from social media data using the opportunity algorithm based on topic modeling and sentiment analysis. The conclusions with further research topics are then presented.

2. Theoretical background

The approach proposed in this paper is based on three theoretical backgrounds: opportunity algorithm, topic modeling, and sentiment analysis. Therefore, this section presents a brief overview of these backgrounds.

2.1. Opportunity algorithm

In the present study, the opportunity algorithm is used to identify the extent to which each product topic is a potential opportunity for improvement from a customer-centered view. The opportunity algorithm, which was proposed along with outcome-driven innovation (ODI) by Ulwick (Ulwick, 2005), is a method used to prioritize unmet needs. The opportunity algorithm uses importance and satisfaction as two factors to compute opportunity on a scale of 0–10 from a customer perspective. The ODI philosophy assumes that an opportunity for innovation exists when a need is important yet is not well satisfied; the opportunity for value creation increases as the customer's need becomes more important and as customer satisfaction decreases. From this perspective, the opportunity algorithm enables firms to find business opportunities for growth by ranking these opportunities in a priority sequence (Yen, Chung, & Tsai, 2007). The needs that are most important, but least satisfied receive the highest priority. The concept of opportunity can therefore be defined as:

$$\text{Opportunity} = \text{Importance} + \text{Max}(\text{Importance} - \text{Satisfaction}, 0) \quad (1)$$

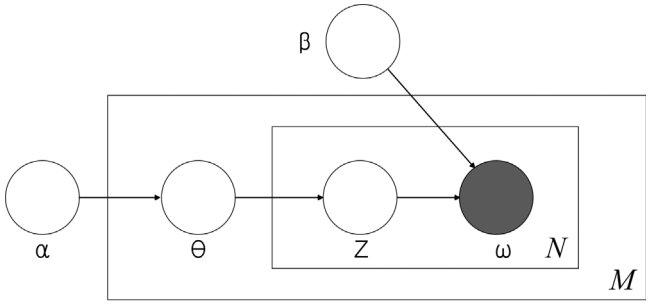
As a simple yet effective heuristic measure that combines the values of importance and satisfaction into a single metric, the opportunity algorithm has been used to define competitive advantage of firms from a resource-based view (Hinterhuber, 2013), to derive customer satisfaction and target management structure (Yen et al., 2007), and to create customer-oriented product portfolios (Helferich, Herzwurm, & Schockert, 2005). In particular, this algorithm was also used to identify the customer needs with the best opportunity from various customer needs and then design optimum products. (Killen, Walker, & Hunt, 2005), and it was applied to develop customer-oriented computer software together with quality function deployment (Andreas Helferich, 2005). Similarly, the present study uses the opportunity algorithm to identify potential opportunities for product topics obtained from social media data from the customer perspective; it quantifies opportunity scores for the product topics and depicts their position using an opportunity landscape map.

2.2. Topic modeling

Since text documents are composed of words, a topic spoken in multiple documents can be expressed by a combination of strongly related words. Each document is considered to belong to multiple topics. In this way, topic modeling is a technique used to infer hidden topics in text documents. Because the topic modeling represents each document as a complex combination of multiple topics and each topic as a complex combination of multiple words, it is also used as a text-mining tool to classify documents based on topic inference results.

In the present study, Latent Dirichlet Allocation (LDA)-based topic modeling is used, which is known to have the highest performance among several topic modeling algorithms when dealing with large-scale documents and interpreting identified latent topics (Chiru, Rebedea, & Ciotec, 2014). The procedure of the LDA model involves three steps as shown in Fig. 2 (Blei, Ng, & Jordan, 2003). LDA assumes the following generative process for a corpus D consisting of M documents, each of length N_i. α is the parameter of the Dirichlet prior to the per-document topic distribution, β is the parameter of the Dirichlet prior to the per-topic word distribution, θ_i is the topic distribution for document i (the sum of θ_i is 1.0), φ_k is the word distribution for topic k, z_{ij} is the topic for the j^{th} word in document i, and w_{ij} is the specific word.

LDA-based topic modeling has been widely used for various purposes, including patent development map generation (Kim, Park, & Yoon, 2016), automatic crime prediction based on the events extracted from social media (Wang, Gerber, & Brown, 2012), web spam-filtering and fraud detection (Bíró, Szabó, & Benczúr, 2008;Xing & Girolami, 2007), and scientific article and website recommendation (Das, Datar, Garg, & Rajaram, 2007 ;Jin, Zhou, & Mobasher, 2005;Krestel, Fankhauser, & Nejdl, 2009; Wang and Blei, 2011). The LDA-based topic modeling has been also used in products planning processes as a building block; it was used to define the



- ✓ Choose $\theta_i \sim Dir(\alpha)$, where $i \in \{1, \ldots, M\}$
- ✓ Choose $\varphi_k \sim Dir(\beta)$, where $k \in \{1, \ldots, K\}$
- ✓ For each word position i, j where $j \in \{1, \ldots, N_i\}$, and $i \in \{1, \ldots, M\}$
  - Choose a topic $z_{ij} \sim Multinomial(\theta_i)$.
  - Choose a word $w_{ij} \sim Multinomial(\varphi_{z_{ij}})$,

Fig. 1. Concept of LDA-based topic modeling (Blei, Ng, & Jordan, 2003).

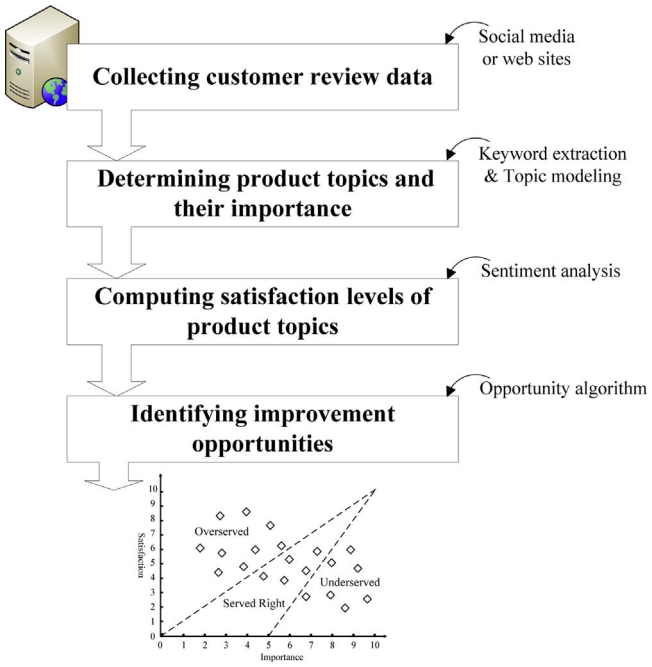


Fig. 2. Overview of the proposed approach.

product portfolios of firms to identify new product opportunities based on patent intelligence (Yoon, Seo, Coh, Song, & Lee, 2017) and to identify sub-technology topics and their competitive intelligence in augmented reality technology for product planning (Jeong & Yoon, 2017). LDA-based topic modeling is a useful technique for latent topic identification from a large corpus; the present study therefore uses it to identify product topics discussed by customers in social media.

2.3. Sentiment analysis

Textual data can be broadly categorized into facts and opinions; facts are objective expressions such as entities and events and their properties, while opinions are subjective expressions that describe people's sentiments, appraisals or feelings (Liu, 2010). As one of the major research fields in natural language processing, sentiment analysis focuses primarily on opinion expressions that convey people's polarity of sentiment. This is because the sentiment analysis involves the task of detecting positive or negative opinions of the speaker who wrote the textual data.

The speaker's sentiment in textual data and the degree of sentiment are generally identified by the lexicon-based, text classification-based, and deep learning-based approaches. The lexicon-based approach uses the predefined dictionaries that define sentiment words and their corresponding sentiment value (e.g. SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010)) and identifies the sentimental orientation of a document based on the semantic orientation of words or phrases in the document (Turney, 2002). The text classification-based approach builds a classifier using labeled instances of texts or sentences and then evaluates the sentimental orientation of a given document based on the classifier constructed by supervised learning (Pang, Lee, & Vaithyanathan, 2002; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). Finally, the deep learning-based approach employs a deep neural network model to analyze the sentiment of speech in the text document (dos Santos & Gatti, 2014;Glorot, Bordes, & Bengio, 2011).

Sentiment analysis has been used to identify customers' opinions for customer-oriented product planning, such as identifying the weaknesses of products from online review data (Zhang, Xu, & Wan, 2012), measuring the customer satisfaction of a mobile service (Kang & Park, 2014), identifying the service quality from online user-generated contents (Duan, Cao, Yu, & Levy, 2013) and defining consumers' brand image from social media (Mostafa, 2013). The present study uses the deep learning-based approach to measure customer's satisfaction within the consumer discussion topics of a given product from social media data. This approach is used because the deep learning-based approach has been effectively used for sentiment analysis of relatively short texts such as Twitter and social media text (dos Santos & Gatti, 2014;Tang, Wei, Qin, Liu, & Zhou, 2014).

3. Proposed methodology

Our approach for product opportunity identification using social media mining is built on three theoretical backgrounds: topic modeling, sentiment analysis, and opportunity algorithm (Fig. 3). First, product topics discussed by customers and the importance value of the product topics are identified by applying topic modeling to large-scale social media data related to the target product under study. Second, the satisfaction value of customers for the product topics are measured based on the application of deep learning-based sentiment analysis of the keywords obtained from social media data. Third, the opportunity score of each product topic is evaluated by the opportunity algorithm composed of the product topics' importance and satisfaction values. As a result, the directions for product improvement are deduced from the major negative sentiment keywords of the product topics with a high opportunity score.

3.1. Step 1: Data gathering and preprocessing

Our approach first involves collecting social media data related to a target product. The material for this approach should be online review data generated by product customers on social media or social websites; this is because our approach identifies major product topics that are currently being discussed by product customers and the satisfaction analysis of the product topics is performed from a customer-centered perspective. In this step, some techniques used to collect large-scale social media data can be adopted, including web crawling techniques and open application programming interfaces (APIs) provided by social media services; for example, some open APIs are available on Twitter (https://dev.twitter.com) and Reddit (https://www.reddit.com/dev/api).

Once a set of customer online reviews of the target product has been prepared, keywords (or key phrases) are extracted from the customer

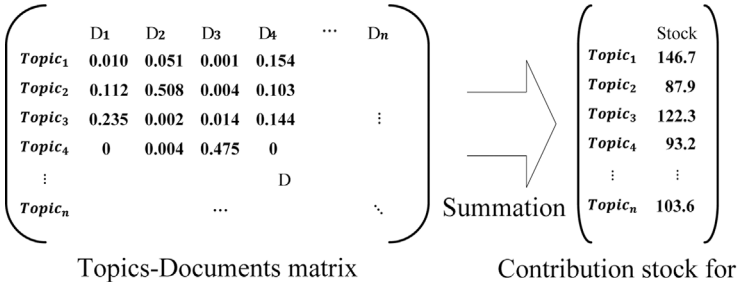


Fig. 3. Concepts of computing the importance of product topics.

online reviews to structure each of the reviews. Although a list of keywords can be obtained by applying text mining to the text of the data, some of the keywords may be irrelevant or too generic for textual analysis. For example, emoticons ('^^', ':-D') and onomatopoeic words ('haha', 'blah'), and irrelevant words ('product', 'process') should be excluded from the keyword list. Finally, each customer online review can be structured as an array of keywords and their frequency that appear in their corresponding online review.

3.2. Step 2: Identifying product topics and their importance

In this step, product topics are defined and the importance value of each product topic is computed using LDA-based topic modeling. LDA-based topic modeling requires three inputs for its execution: corpus, word dictionary, and the number of topics. In this step, the corpus and the word dictionary become the customer online reviews and the keyword list obtained from the previous step, respectively. Next, to determine the appropriate number of topics for LDA-based topic modeling, the elbow method is adopted, in which an optimal number of topics is determined by calculating the average cosine similarity between all pairs of topic-word distribution vectors produced by topic modeling (Wang, Liu, Ding, Liu, & Xu, 2014).

Product topics outputted by LDA-based topic modeling become the major subjects by which customers are discussing the target product because the product topics are derived from the customer-generated online review data. In addition, the degree of importance of the product topics can be measured by calculating the degree to which each product topic is mentioned by the users. Accordingly, the sum of the contribution probabilities of each product topic to all customer reviews indicates the importance of the product topic within the collected corpus (Fig. 3). Then, the importance of product topics is normalized on a scale of 0–10 to generate the values of importance dimension for the opportunity algorithm (Eq. (2)).

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

$$Importance_i = 10 \times \frac{CS_i - CS_{Min}}{CS_{Max} - CS_{Min}} \quad (2)$$

3.3. Step 3: Computing the satisfaction level of product topics

The aim of this step is to measure the customer satisfaction level of each product topic. As previously mentioned in Section 2.3, several techniques such as lexicon-based, text classification-based, and deep learning-based methods can be used to identify emotional orientations of terms or documents spoken. First, the average sentiment score for each keyword is defined using sentiment analysis and a keyword sentiment vector, or an array composed of all keywords and their average sentiment score, is then generated. Next, a sentiment-weighted topic-keyword matrix can be constructed by multiplying the keyword sentiment vector with the topic-keyword matrix outputted by topic modeling in the previous step.

Once the sentiment-weighted topic-keyword matrix is obtained, the satisfaction degree of each product topic is computed by summing the sentiment-weighted values of keywords constituting their corresponding product topic (Fig. 4). For the application of the opportunity algorithm in the next step, in this step the satisfaction degree of product topics is transformed to a scale of 0–10 (Eq. (3)).

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

$$Satisfaction_i = 10 \times \frac{SS_i - SS_{Min}}{SS_{Max} - SS_{Min}} \quad (3)$$

3.4. Step 4: Identifying product opportunities using the opportunity algorithm

In this step, by applying the opportunity algorithm to the product topics' importance and satisfaction scores, the opportunity potential of each product topic is evaluated and an opportunity landscape map is generated. According to Eq. (1), the product topics that are most important but least satisfied have the highest opportunity; such product topics could be used to determine the subject for further product improvement. In addition, the opportunity landscape map can be drawn from the importance and satisfaction values. This map can be divided into three areas: served-right, over-served, and underserved (Fig. 1). Needs in the served-right area are considered appropriately satisfied, needs in the over-served area are considered excessively satisfied, and needs in the underserved area are considered less satisfied, compared to the importance of the needs. Therefore, based on the opportunity algorithm, underserved needs are understood as innovation opportunities.

Once the product topics with a high opportunity score have been identified, development directions can be set using the sentiment-weighted topic-keyword matrix. In this matrix, each product topic has its contributing keywords and their sentiment weighted values. Therefore, primary directions for product opportunities of each product topic can be formulated towards increasing the satisfaction or decreasing the dissatisfaction of the keywords with a high sentiment weighting within the product topic.

4. Case study: samsung galaxy note 5

In this section, the process of the proposed approach is illustrated using social media data related to the Samsung Galaxy Note 5 (SGN5). This smartphone is a multifunctional product as it provides customers with generic functions, including phone calls, wireless internet, and photography, as well as device-specific functions, including an S-pen, wireless charging, and Samsung payment. In addition, because SGN5 was previously sold worldwide, large-scale product customers who have used this smartphone are actively discussing this product and exchanging their opinions via social media. Therefore, we considered that the multi-functionality and customer size of SGN5 can be used to clearly explain our product opportunity analysis process.

4.1. Data gathering and preprocessing

Social media data can be gathered from various forums such as Twitter, Facebook, Instagram, and Reddit. The characteristics of these forums are relatively similar. For example, Twitter users can post a short message limited to 140 characters and they can discuss various topics using hashtags, words that begin with the hashtag symbol (#). Among various social medias, Reddit (https://www.reddit.com) was used for this case study. Reddit entries are organized into subreddits that are areas of interest. Each subreddit is a type of forum in which users of Reddit can write posts on any subreddit that include their interest topics and can write their opinion on comments of posts. Generally, the acquired data, from social media such as Twitter and Facebook, are limited, since the data can only be collected through keywords or hashtags selected by the analyst. On the other hand, the data in subreddits include both direct and indirect opinions of a product, regardless of the occurrence of product keywords. Thus, Reddit data is used for this case study (Fig. 5).

A total of 23,614 documents, including 2255 posts and 21,359 comments, were gathered from the subreddit of SGN5 (https://www.reddit.com/r/galaxynote5) and the quantitative trend of the data is shown in Fig. 6. As seen in the figure, they were written between 07 Nov 2014 (GMT) and 31 Jan 2016 (GMT), and the number of posts increased explosively around the release date (21 Aug 2015). The number of total posts and comments before the release date is 1266,

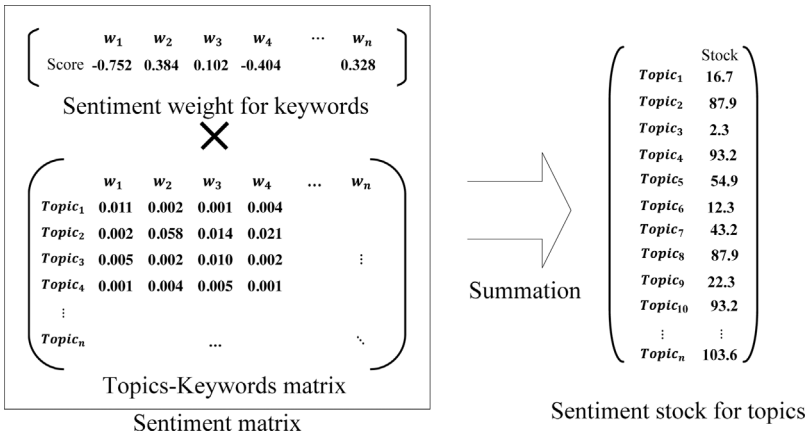


Fig. 4. Concepts of computing the satisfaction of product topics.

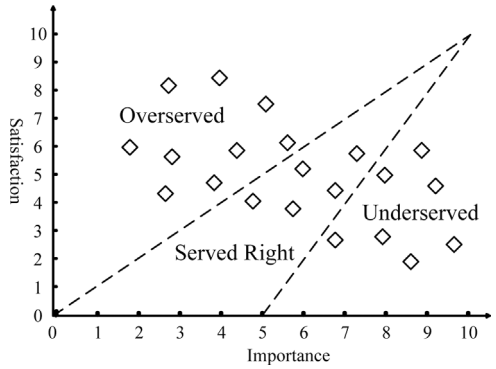


Fig. 5. Schematic of opportunity landscape maps.

implying that 4.41 posts and comments were written in one day. On the other hand, 22,345 posts and comments were written between 21 Aug 2015 and 31 Jan 2016. On average, the number of posts and comments were written in one day is 136.25, implying an increase of 31 times compared to before the SGN5 release date.

Although all documents were originated in the subreddit of SGN5, not all documents were related to SGN5, because advertisements and meaningless documents such as "Solar Charger with Big LED Light, DT® 38,000mAh Portable Dual-Port Solar Charger" and "Imgur[deleted]" are included in the gathered documents. Accordingly, these noise documents should be eliminated. In this case study, noise documents are defined as those with no keywords that relate to 'SGN5'.

Keyword:s were extracted from all documents by the Rapid automatic keyword extraction algorithm (RAKE). RAKE is an unsupervised; domain-independent; and language-independent method for extracting keywords from individual documents (Rose, Engel, Cramer, & Cowley, 2010). It has the advantage of being able to extract key phrases (compound words). An initial set of 32,637 keywords were obtained; but it contained many meaningless keywords such as webpage links ('www.amazon.com'; 'www.apkmirror.com') and colloquial words ('WoW'; 'Yeah'). Noise keywords were selected using 3 steps. First; keywords that have an occurrence frequency of less than 1 in each document were removed. Second; keywords that have a document frequency (DF) of less than 1 were removed because they are local words that do not effect LDA; the occurrence-based analysis. Lastly; keywords unrelated to the SGN5 among the remaining keywords were manually removed. Through this process; 3539 keywords were confirmed for analysis. In addition; 12,491 documents were found not to contain these confirmed keywords; so they were considered as noise documents and excluded from the document set for analysis. Finally; 3539 keywords and 11,123 documents were confirmed for this case study.

4.2. Identifying product topics

The keyword frequency matrix with confirmed keywords and documents was constructed for LDA. Some software is available that supports LDA, such as gensim (a python package), tm (an R package) and Netminer (commercial software). We used Netminer to execute LDA topic modeling because it is easy to use and various options and parameters can be selected. Each row of the matrix refers to the individual documents identified with 'Document ID', each column refers

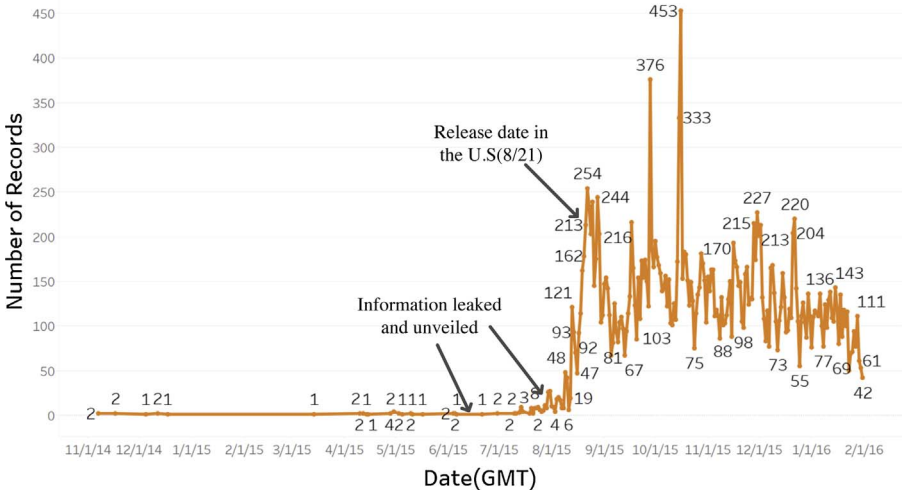


Fig. 6. Quantitative trend of SGN5 subreddit.

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

to keywords (Table 1), and each cell refers to the occurrence frequency of the keyword in the document. The LDA uses this matrix as the input and factorizes the matrix into two matrices: a Documents-Topics matrix (Table 2) and a Topics-Keywords matrix (Table 3).

In addition, the LDA topic modeling needs the number of topics as another major input. This input can be defined by various approaches such as a literature study and similarity-based approach. In this case study, the similarity-based approach was used to define the number of topics of the LDA. The main idea of this approach is that the number that best separates the topics is the optimal number of topics (Wang et al., 2014). In this same vein, the lowest point of similarity between overall pairs of topics is an optimal number of topics. The similarity-based approach was used in this study and cosine similarity was used as a similarity measure index. Cosine similarity is a widely used similarity measure. Two vectors, assumed as A and B, have an angle (θ) in a vector space. Accordingly, the value of cos(θ) between A and B can be calculated (Eq. (4)). If two vectors are the same, their angle is 0° and the cosine similarity is 1. On the other hand, if the angle between A and B is 90°, the cosine similarity of these two vectors is 1. According to similarity-based analysis, the number of topics for SGN5 data is defined as 65 because the inflection point is 65 (average cosine similarities of pairs of topics = 0.04554) (Fig. 7).

$$\text{Cosinesimilarity(A, B)} = \cos(\theta) = \frac{A \cdot B}{AB} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4)$$

Sixty-five topics from LDA were named by numbering them, such as ‘Topic-1′, ‘Topic-2′, and ‘Topic-3′. The name does not give any information about the topic. Thus, the real name of topics was manually defined by referring to the documents and keywords that have a top contribution to the topic. For example, ‘Topic-10′ was named ‘Expandability’ because major keywords and their contributions of ‘Topic-10′ are ‘SD (0.130)’, ‘SD-card (0.094)’, ‘removable battery (0.049)’, ‘MicroSD (0.036)’, and ‘IR blaster (0.016)’. Furthermore, several social data that have the highest contribution for ‘Topic-10′ include customer comments such as, “I don't understand. If the Note 5 was even the slightest bit better than the Note 4 and had a microSD card slot/removable battery, it'd sell like hotcakes. My hopes and dreams (0.240)” and “so you are disappointed that the new Note doesn’t have a removable battery and no SD Card slot so you consider... another phone with no removable battery and no SD Card. Seems legit. (0.232)”. Other 64 topics were named in the same way such as ‘Design’, ‘Fingerprint’, ‘Samsung Pay’. Some topics and their major keywords and representative social media data are explained in Table 4. Among all topics, two topics have a similar name, ‘Detect pen1′ and ‘Detect pen2′, although they slightly differ, whereby ‘Detect pen1′ is related to detecting the rotation and position on the screen of an S-pen, while ‘Detect pen2′ is related to recognizing the insertion and ejection of the S-pen.

Among the 65 topics, six relate to the battery and its charge, including ‘Battery life’, ‘Battery usage’, ‘Battery drain’, ‘Fast charge’, ‘Charger’, and ‘Wireless charge’. It appears that these topics were mentioned by many customers. In addition, the new or improved features of a previous product, the Samsung Galaxy Note 4, were captured in the topic analysis as 6 topics, including ‘Samsung Pay’, ‘Pen out’, ‘Expandability (including external memory card and replaceable battery)’, ‘Wireless charge’, ‘Screen off memo’, and ‘Samsung Pay on ATM’. Customers also participated in many discussions on these new topics because they are some of the most attractive features of the improved product. Other product topic clusters that are most often mentioned in SGN5 are software-related topics including OS, Android, and Apps, because they were also captured in 13 topics, including ‘Software update’, ‘Music App’, ‘Multi-window’, ‘Touch wiz’, ‘Widget’, ‘Custom Rom’, ‘OS upgrade’, ‘OTA’, ‘Multi-tasking’, ‘Default application’, ‘New OS feature’, ‘Play store’, and ‘Screen off memo’. This demonstrates that software is as important as hardware in the current smartphone market.

4.3. Computing importance and satisfaction degree of product topics

In order to compute the importance and satisfaction of topics, the two matrices from LDA, the Documents-Topics matrix and Topics-Keywords matrix, were used to compute the degrees of importance and satisfaction. The degree of importance of product topics was computed from Eq. 2 with the Documents-Topics matrix. The contribution stock of topics was calculated from the sum of topic portions in each document. According to the analysis, the highest contribution stock is 202.2198 (Topic: ‘Samsung Pay’) and the lowest contribution stock is 165.9974 (Topic: ‘Accessory’). The average and standard deviation of contribution stocks are 171.1228 and 6.58, respectively. In addition, the number of topics that have a higher contribution stock than the average value is only 20 among the 65 topics. To summarize, several abnormalities occur, such as ‘Samsung Pay (202.2198)’, ‘Fast charge (191.5701)’ and ‘Software update (182.7272)’. They also have a higher importance (Table 5) and they may have a higher rank in opportunity analysis.

Table 2
Part of Documents-Topics matrix of SGN5 .

Document ID	Topic-1	Topic-2	Topic-3	Topic-4	Topic-5	Topic-6	Topic-7	Topic-8	Topic-9
2	0.0159	0.0137	0.0137	0.0141	0.0139	0.0139	0.0149	0.0141	0.0139
3	0.0147	0.0145	0.0147	0.0147	0.0152	0.0145	0.0147	0.0152	0.0145
4	0.0135	0.0135	0.0135	0.0135	0.0145	0.0135	0.0149	0.0139	0.0139
5	0.0160	0.0118	0.0118	0.0128	0.0139	0.0118	0.0157	0.0125	0.0118
6	0.0185	0.0124	0.0127	0.0127	0.0129	0.0127	0.0124	0.0124	0.0136
7	0.0337	0.0136	0.0051	0.0102	0.0100	0.0059	0.0093	0.0054	0.0273
8	0.0691	0.0054	0.0054	0.0187	0.0073	0.0052	0.0131	0.0058	0.0199
10	0.0159	0.0148	0.0152	0.0154	0.0154	0.0152	0.0150	0.0148	0.0148
11	0.0120	0.0120	0.2301	0.0120	0.0120	0.0120	0.0120	0.0120	0.0120
12	0.0116	0.0112	0.0108	0.0121	0.0133	0.0108	0.0149	0.0113	0.0116
13	0.0145	0.0142	0.0142	0.0151	0.0145	0.0147	0.0142	0.0145	0.0142
15	0.0106	0.0103	0.0099	0.0102	0.0105	0.0098	0.0106	0.0100	0.0105

Table 3
Part of Topics-Keywords matrix of SGN5 .

Topics	4G	4GB	Exynos	OS	RAM	64GB	battery	BT	car
Topic-1	0.0001	0.0001	0.0002	0.0095	0.0012	0.0001	0.0007	0.0332	0.0006
Topic-2	0.0001	0.0001	0.0001	0.0031	0.0001	0.0001	0.0002	0.0001	0.0005
Topic-3	0.0001	0.0001	0.0001	0.0007	0.0002	0.0001	0.0001	0.0008	0.0001
Topic-4	0.0002	0.0001	0.0002	0.0004	0.0002	0.0001	0.0491	0.0022	0.0003
Topic-5	0.0001	0.0001	0.0002	0.1254	0.0002	0.0001	0.0037	0.0002	0.0103
Topic-6	0.0001	0.0001	0.0001	0.0023	0.0002	0.0001	0.0002	0.0004	0.0001
Topic-7	0.0001	0.0001	0.0003	0.1885	0.0001	0.0001	0.0002	0.0002	0.0005
Topic-8	0.0001	0.0001	0.0003	0.0204	0.0009	0.0001	0.0001	0.0005	0.0001
Topic-9	0.0001	0.0001	0.0003	0.0027	0.0003	0.0001	0.0004	0.0011	0.0002

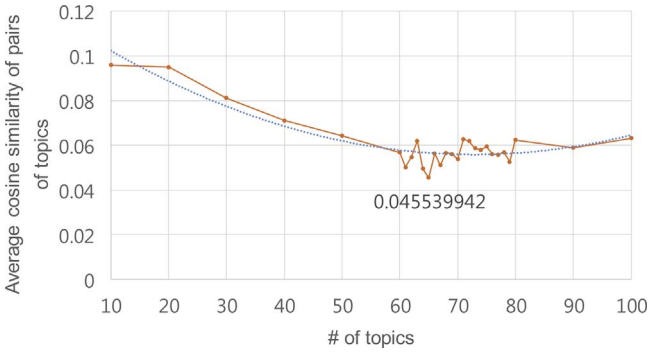


Fig. 7. Average cosine similarities of pairs of SGN5 topics.

Table 5
Part of a sentiment weight of keywords.

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 accessories	0.2027 −0.0810	auto-focus Battery life	0.0334 −0.0457	charger OS systems	−0.1597 0

Two further steps are needed for the degree of satisfaction. To compute the satisfaction degree of each topic, a deep learning-based sentiment analysis was conducted. While it could have been implemented using polarity datasets such as movie review data on the web (http://www.cs.cornell.edu/people/pabo/movie-review-data/) (Pang et al., 2002), we use AlchemyAPI included in IBM's Watson platform, which has many functions related to natural language processing, not only keyword-level sentiment analysis but also documents sentiment analysis, entities extract, taxonomy classification, and others. The time span for implementing and testing the sentiment classifier can be reduced by using AlchemyAPI and accuracy can be guaranteed because it uses large-scale data to train the classifier.

According to the sentiment analysis, 3539 keywords occur a total of 105,483 times in 11,123 document sets by different sentiments. For example, the sentiment score for 'battery' in the phrase, "I will blame the phone on that because only on an android do you have rogue apps draining battery. That does not happen on iPhone" is −0.9013, but the sentiment score is 0.8790 for the phrase, "My battery is completely fine and pretty much exactly what you'd expect". Thus, the average sentiment score of each keyword was used as a sentiment weight for keywords (Table 5). Then, the sentiment matrix is created by multiplying

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."

Table 6
Degrees of importance and satisfaction of product topics.

Topic	Importance	Satisfaction	Topic	Importance	Satisfaction
Samsung Pay	10.0000	7.8185	Battery usage	0.6722	3.6543
Fast charge	7.0599	5.4720	Battery drain	0.6077	2.8013
Detect pen2	4.3805	0.1384	OS upgrade	0.6013	7.7676
Pen out	4.6130	0.6877	Location	0.5573	4.5419
Battery life	4.2233	1.7818	OTA	0.4712	4.5236
Expandability	3.6873	2.1963	Warranty & Repair	0.4699	6.5632
Software update	4.6186	4.9924	Galaxy note5	0.4645	6.7488
Detect pen1	2.2850	0.0000	Screen resolution	0.4159	4.5756
Charger	4.3514	4.4286	etc	0.4057	3.5863
Screen glass	3.9796	4.8418	UX	0.4033	4.5284
Screenshot	3.0895	4.3547	E-mail	0.3966	4.6781
Lock screen	2.5322	4.0059	Hardware Spec.	0.3668	7.0357
Write on screen	1.4050	0.3948	Calling	0.3514	3.5663
Wireless charge	2.2300	6.1661	Optimization	0.3409	4.7985
Wi-Fi	2.1173	3.9489	Edge display	0.3276	6.2638
Screen(AMOLED)	1.9574	4.0696	Multi-tasking	0.3065	5.5286
Emoji	1.6663	2.7311	Default application	0.2974	5.3414
Music App	1.5975	4.7520	Data Transfer	0.2818	7.0303
Camera	1.5629	8.1430	Case	0.2805	4.9050
Multi windows	1.5081	3.7140	Game	0.2768	5.2511
Hand write	1.4836	5.6843	New OS feature	0.2706	7.7689
Touch wiz	1.3709	10.0000	Stylus	0.2666	5.8965
Widgets	1.1853	5.2287	SIM	0.2406	5.1276
Fingerprint	1.0090	4.2611	Play store	0.2351	5.3711
Custom Rom	0.9580	5.0486	Design	0.2306	5.3133
Internal storage	0.9452	6.3186	Screen off memo	0.2182	4.4906
Icon & wallpaper	0.9126	5.2325	Samsung Pay on ATM	0.2052	6.3513
Physical buttons	0.8501	2.5684	Sound	0.1751	4.7423
Device connect	0.8404	5.2066	Material	0.1687	5.3335
Video record	0.7990	5.0560	Google Play	0.1680	6.6631
Theme	0.7744	5.6222	Hardware performance	0.0822	6.5822
Network dropped	0.7387	1.3132	Accessory	0.0000	6.3538
SMS	0.6854	1.1920	–	–	–
–	–	–	**Average**	**1.4150**	**4.7854**

the sentiment weight for keywords by each row of the Topics-Keywords matrix. The sentiment stock of topics was then calculated by the sum of all values in each row of the sentiment matrix. Finally, the degree of satisfaction of each topic was computed by Eq. (3). Consequently, the highest sentiment stock is 0.1093 (Topic: 'Touch wiz') and the lowest is −0.2918 (Topic: 'Detect pen1′). The average and standard deviations of sentiment stocks are −0.0999 and 0.0791, respectively. In addition, the number of topics that have a higher sentiment stock than the average value is 36 among the 65 topics.

Next, for opportunity analysis, contribution stocks (importance) and sentiment stocks (satisfaction) for all product topics were respectively normalized by their maximum and minimum values. The degrees of importance and satisfaction after normalizing for all topics are shown in Table 6.

4.4. Identifying product opportunities

The opportunity score of all topics was identified by Eq. (1) using the degrees of importance and satisfaction (Table 7) and the opportunity landscape map of SGN5 was then drawn (Fig. 8). According to the landscape map, the number of under-served topics that also have a higher opportunity score is 7 and these under-served topics include 'Samsung Pay', 'Fast charge', 'Battery life', 'Expandability', 'Pen out', 'Detect pen1′, and 'Detect pen2′. They have a lower degree of satisfaction than the degree of importance. First, 'Samsung Pay' is a new feature, first released in SGN5, and does not exist in any previous smartphones. Thus, it seems that this topic was often mentioned by customers and it also had a higher desire and expectation. Second, 'Fast charge' and 'Battery life' are ultimately related to longer usage time. Although various solutions have been applied to this topic such as higher-voltage charge, wireless charge, new material (lithium-polymer), and software techniques, customers seems to still have a number of other desires for these product topics. Third, 'Expandability' is an expected topic that has a higher opportunity. Major keywords of 'Expandability' are an 'external memory card' and 'replaceable battery' and the 'replaceable battery' is also related to 'longer usage time'. Lastly, the pen-related topics, 'Pen out', 'Detect pen1′, and 'Detect pen2′ are other under-served topics. The stylus pen, called S-pen, is the key feature of the Galaxy Note series that is not available on other smartphones.

Although improvement opportunities were analyzed using the opportunity landscape map, specific improvement directions are still needed. To address this need, the sentiment matrix can be used to devise a development strategy. In other words, a product development strategy can be made by referencing major keywords of topics in the sentiment matrix. Table 8 shows the major positive and negative keywords of six topics that have a higher opportunity.

The topic *'Samsung Pay'* has the highest opportunity in the topics of SGN5 and major negative keywords of 'Samsung Pay' are related to near field communication (NFC) and competitive services such as 'Apple pay' and 'Android pay', which support NFC payment, although 'Samsung Pay' does not yet support NFC payment. However, SGN5 customers seem to have had a positive experience with the Samsung pay and Samsung pay promotion events. Thus, 'Samsung Pay' needs to support NFC and increase the detection rate. The *'fast wireless charging'* is another feature first released in SGN5. It was implemented by increasing the charge power by 1.5 times more than existing wireless charging: from 5.0 V 2.0A (10Wh) to 9.0 V 1.67A (15Wh). According to the analysis, it seems customers are satisfied with the functions of a quick charge and fast wireless charging but are unsatisfied with the wireless charger due to its problems in connection with the charging pauses. Thus, the detection rate of charge of the fast wireless charger needs to be increased. Actually, the most recent charger (Model name: EP-NG930) has been released by adding one extra charge coil to

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**

increase the detection area. The topic of *'Detect pen2'* concerns the detection of the eject and insert of an S-pen. In particular, the new eject mechanism, called push-to-eject, was applied in SGN5; however, it causes a flaw whereby the SGN5 breaks when the stylus is inserted backwards. Thus, the major negative keywords show this flaw: 'S-Pen sensor', 'S-Pen Backwards', and 'spring mechanism'. However, because this flaw was resolved in the latest revision of SGN5, this topic does not have a problem. For the topic of *'Battery life'*, although the screen-on time satisfies some customers, other customers are unsatisfied with the battery usage time of SGN5. This is because some software techniques used to increase the usage time are captured in this topic, such as 'Package Disable', 'background service', and 'unnecessary process'. The *'Package Disable'* is a function of an application 'Package Disabler', which can disable the default application. The other two keywords, 'background service' and 'unnecessary process', are techniques used to kill unnecessary processes and services. In this topic, both hardware and software need improvement to increase usage time. The topic of *'Expandability'* includes not only external memory and replaceable battery but also third party equipment. Although many customers oppose the removal of the MicroSD slot, they seem to find a solution such as a cloud storage. This phenomenon of finding an alternative solution also appeared in the replaceable battery. Many customers are satisfied with an external battery, which appeared in major positive keywords such as 'battery-pack' and 'portable power banks'. The 'Software update' topic has a positive image for customers and can be found through major positive keywords such as 'optimization'. In fact, the keyword 'optimization' occurs nine times in the total number of documents, e.g. "Just take a moment to appreciate the battery since the app optimization update". However, the customers have shown dissatisfaction with the *'Software update'* topic, as indicated by the keyword 'factory reset'. After the update release, some customers recommended the factory reset or the reinstallation of the whole system after a problem occurred when completely removing the trashed items of the previous system. Actually, problems can sometimes be solved by a factory reset, as expressed by the phrase, "I reset my phone and it worked. My phone was doing a bunch of weird stuff so I decided to do a factory reset from the settings, and now it worked. So weird!". However, the factory reset is difficult for general customers. Another problem is overheating when the software updates and with battery drain. This problem needs to be addressed when updating the software.

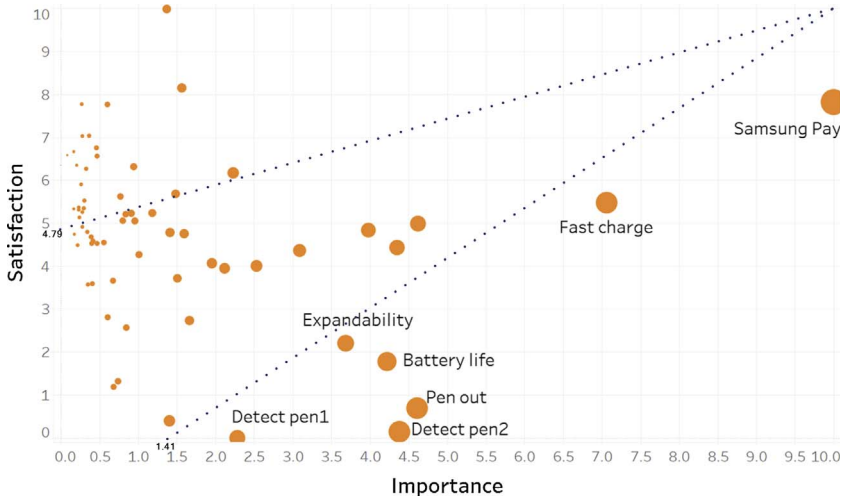


Fig. 8. Opportunity landscape map of SGN5 (Circle size = Opportunity score).

Table 8
Part of five negative and five positive keywords for product topics.

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

5. Concluding remarks

A social media mining approach was proposed in this study for identification of product development opportunities. As the building blocks of the approach, this study uses LDA-based topic modeling, sentiment analysis, and opportunity algorithm. In terms of the specific steps of the approach, each product topic from customers' perspective was defined by LDA using customer-generated social media data. The degrees of importance and satisfaction of each topic were then computed. The importance of the topic is computed based on the concept of the contribution stock and the satisfaction of topic is computed based on the concept of the sentiment stock using sentiment matrix. Finally, the opportunity value and improvement direction of product topics are identified from a customer-centered view using the opportunity algorithm. The functionality of the approach was demonstrated herein using the SGN5 data on Reddit, one of the major social media in the United States, between 07 Nov 2014 (GMT) and 31 Jan 2016 (GMT). The development directions on the top six topics of the SGN5 were found through this case study. The proposed domain-independent approach contributes to the exploration of new product opportunities across various domains, including not only products but also services, product-service systems, and software, using the social media data related to the target on the web. It thereby assists product planners in the design step to identify new or improved products by capturing uncaught opportunities.

We expect that this study will make both academic and industrial contributions to relevant fields. From an academic perspective, the proposed approach defines customer attractive topics and quantifies product development opportunities of topics using customer-generated social media data. Some limitations of prior studies were revealed. For example, they focused only on polarity detection while neglecting potential opportunities and term-level analysis was conducted in which it is difficult to define semantic topics mentioned by customers. On the other hand, in our approach, a topic analysis is conducted and latent opportunities are identified. In addition, while most prior studies do not provide a guideline for product development, our approach can suggest a guideline for each topic using the sentiment matrix. Our approach can also be used to monitor the trends of customer needs if the publication time of customer review data is considered. From an industrial perspective, our approach can be implemented as a software system for firms. Because recent customer needs are more dynamic and business environments have become globalized, it is important that they rapidly track and deal with evolving customer needs in order to retain their competitive position. When this is the case, our proposed approach will be an efficient aid, which can early identify and prioritize dynamic customer needs.

Despite the contributions made by this study, further works still need to be completed. First, in the proposed approach a relative degree of both importance and satisfaction is applied through normalizing. This is because determining the absolute value that corresponds to a scale of 0–10 is difficult. Accordingly, the standard degrees of importance and satisfaction need to be defined for future research. Second, over-served needs in the opportunity landscape map need to be more examined because they can be a clue to achieved disruptive innovation (Silverstein et al., 2013). Therefore, in future works, the product opportunities for disruptive innovation can be captured. Finally, our approach was applied to one example target product, but it has the potential to be applicable to various domains, such as services and product-service systems. Therefore, application studies in different domains will be conducted in further works.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2015R1A1A1A05027889).

References

Bíró, I., Szabó, J., & Benczúr, A. A. (2008). Latent dirichlet allocation in web spam filtering. *In Proceedings of the 4th international workshop on Adversarial information retrieval on the web ACM,* 29–32.
Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *In LREC, 10*, 2200–2204.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.
Brooks, S. (2015). Does personal social media usage affect efficiency and well-being? *Computers in Human Behavior, 46*, 26–37.
Chiru, C.-G., Rebedea, T., & Ciotec, S. (2014). Comparison between LSA-LDA-Lexical chains. *In WEBIST, 2*, 255–262.
Das, A. S., Datar, M., Garg, A., & Rajaram, S. (2007). Google news personalization: scalable online collaborative filtering. *In proceedings of the 16th international conference on world wide web,* 271–280 [ACM].
Duan, W., Cao, Q., Yu, Y., & Levy, S. (2013). Mining online user-generated content: using

sentiment analysis technique to study hotel service quality. *In system sciences (HICSS), 2013 46th hawaii international conference on IEEE,* 3119–3128.
Geum, Y., Lee, H., Lee, Y., & Park, Y. (2015). Development of data-driven technology roadmap considering dependency: An ARM-based technology roadmapping. *Technological Forecasting and Social Change, 91*, 264–279.
Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: a deep learning approach. *In proceedings of the 28th international conference on machine learning (ICML-11),* 513–520.
Griffin, A., & Hauser, J. R. (1993). The voice of the customer. *Marketing Science, 12*, 1–27.
Helferich, A., Herzwurm, G., & Schockert, S. (2005). Mass customization of enterprise applications: Creating customer-Oriented product portfolios instead of single systems. *In proceedings of the 3rd interdisciplinary world congress on mass customization*.
Helferich, A. (2005). Developping customer-oriented enterprise applications using software product lines and quality function deployment. *In proceedings of the 2nd international software product lines young researchers workshop (SPLYR)*.
Hinterhuber, A. (2013). Can competitive advantage be predicted?: Towards a predictive definition of competitive advantage in the resource-based view of the firm. *Management Decision, 51*, 795–812.
Competitive intelligence analysis of augmented reality technology using patent information. *Sustainability9*, 497.
Jin, X., Zhou, Y., & Mobasher, B. (2005). A maximum entropy web recommendation system: Combining collaborative and content features. *In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining ACM,* 612–617.
Kang, D., & Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications, 41*, 1041–1050.
Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media?: Get serious! Understanding the functional building blocks of social media. *Business Horizons, 54*, 241–251.
Killen, C. P., Walker, M., & Hunt, R. A. (2005). Strategic planning using QFD. *International Journal of Quality and Reliability Management, 22*, 17–29.
Kim, M., Park, Y., & Yoon, J. (2016). Generating patent development maps for technology monitoring using semantic patent-topic analysis. *Computers and Industrial Engineering*.
Krestel, R., Fankhauser, P., & Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. *In Proceedings of the third ACM conference on Recommender systems ACM,* 61–68.
Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing, 2*, 627–666.
Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*, 4241–4251.
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing: 10*, (pp. 79–86).
Park, H., & Yoon, J. (2015). A chance discovery-based approach for new product–service system (PSS) concepts. *Service Business, 9*, 115–135.
Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining,* 1–20.
Silverstein, D., Samuel, P., & DeCarlo, N. (2013). *The innovator's toolkit: 50+ techniques for predictable and sustainable organic growth.* John Wiley & Sons.
Sverdlov, G. (2012). Global social technographics update 2011: US and EU mature, emerging markets show lots of activity. *Verkregen Op,* 19.
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*, 267–307.
Tang, D., Wei, F., Qin, B., Liu, T., & Zhou, M. (2014). Coooolll: A deep learning system for Twitter sentiment classification. *In proceedings of the 8th international workshop on semantic evaluation (SemEval 2014),* 208–212.
Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424).
Ulwick, A. W. (2005). *What customers want: Using outcome-driven innovation to create breakthrough products and services*, *Vol. 71408673*. New York: McGraw-Hill.
Van Kleef, E., van Trijp, H. C., & Luning, P. (2005). Consumer research in the early stages of new product development: A critical review of methods and techniques. *Food Quality and Preference, 16*, 181–201.
Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining ACM,* 448–456.
Wang, B., Liu, S., Ding, K., Liu, Z., & Xu, J. (2014). Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: A case study in LTE technology. *Scientometrics, 101*, 685–704.
Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. *In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 231–238).
Wang, X., Yu, C., & Wei, Y. (2012). Social media peer communication and impacts on purchase intentions: A consumer socialization framework. *Journal of Interactive Marketing, 26*, 198–208.
Xing, D., & Girolami, M. (2007). Employing Latent Dirichlet Allocation for fraud detection in telecommunications. *Pattern Recognition Letters, 28*, 1727–1734.
Yen, T.-M., Chung, Y.-C., & Tsai, C.-H. (2007). Business opportunity algorithm for ISO 9001: 2000 Customer satisfaction management structure. *Research Journal of Business Management, 1*, 1–10.
Yoon, J., Seo, W., Coh, B.-Y., Song, I., & Lee, J.-M. (2017). Identifying product opportunities using collaborative filtering-based patent analysis. *Computers and Industrial Engineering, 107*, 376–387.
Zhang, W., Xu, H., & Wan, W. (2012). Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications, 39*, 10283–10291.
dos Santos, C. N., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. *In COLING,* 69–78.