Automated Discovery of Lead Users and Latent Product Features by Mining Large Scale Social Media Networks

Suppawong Tuarob
Computer Science and Engineering,
Industrial and Manufacturing Engineering,
The Pennsylvania State University,
University Park, PA 16802
e-mail: suppawong@psu.edu

Conrad S. Tucker
Engineering Design and Industrial
and Manufacturing Engineering,
Computer Science and Engineering,
The Pennsylvania State University,
University Park, PA 16802
e-mail: ctucker4@psu.edu

Lead users play a vital role in next generation product development, as they help designers discover relevant product feature preferences months or even years before they are desired by the general customer base. Existing design methodologies proposed to extract lead user preferences are typically constrained by temporal, geographic, size, and heterogeneity limitations. To mitigate these challenges, the authors of this work propose a set of mathematical models that mine social media networks for lead users and the product features that they express relating to specific products. The authors hypothesize that: (i) lead users are discoverable from large scale social media networks and (ii) product feature preferences, mined from lead user social media data, represent product features that do not currently exist in product offerings but will be desired in future product launches. An automated approach to lead user product feature identification is proposed to identify latent features (product features unknown to the public) from social media data. These latent features then serve as the key to discovering innovative users from the ever increasing pool of social media users. The authors collect 2.1×10^9 social media messages in the United States during a period of 31 months (from March 2011 to September 2013) in order to determine whether lead user preferences are discoverable and relevant to next generation cell phone designs. [DOI: 10.1115/1.4030049]

1 Introduction

In highly competitive market segments, companies must continually search for next generation product innovations in order to avoid competing solely on price [1]. Multiple research studies have demonstrated the importance of including customers in the product innovation process [2–5]. Recently, an increasing number of companies have altered their product innovation paradigms by making customers the center of product development process, rather than perceiving them simply as the end consumers [6]. More formally, the term *lead user* is defined by von Hippel as customers who [7,8]

(1) face needs that will be general in a marketplace-but face them months or years before the bulk of that marketplace encounters them.
(2) are positioned to benefit significantly by obtaining a solution to those needs.

Consistent with the literature, we define a *lead user* in this work as a consumer of a product that faces needs unknown to the public. Research findings indicate that 10–40% of users have augmented existing products to address latent needs unknown to designers or existing customers [9]. Lead user needs are often converted into potential product development ideas and subsequently incorporated into next generation products. For example, 3M assembled a team of lead users which included a veterinarian surgeon, a makeup artist, doctors from developing countries, and military medics [10]. The recruited lead users then brain-stormed their ideas in a two-and-half day workshop. The successful implementation of 3M's lead user initiative resulted in three product lines (i.e., Economy, Skin Doctor, and Armor lines) that generated eight times more profit than if they had employed traditional product development methods of customer needs extraction [11]. However, a major drawback of such customer-driven paradigms is that only a fraction of customers have the potential to generate innovative ideas useful for next generation product design. This emphasizes the importance of accurately and efficiently selecting lead users from a large pool of potential customers.

Given the abundance of large scale, publicly available data, an interesting research direction worthy of scientific pursuit is whether automated methods that discover lead users and their preferences are viable in the age of social media networks. Society generates more than 2.5 quintillion (10^{18}) bytes of data each day [12]. A substantial amount of this data is generated through social media services such as *Twitter*, *Facebook*, and *Google* that process anywhere between 12 terabytes (10^{12}) to 20 petabytes (10^{15}) of data each day [13]. Social media allows its users to exchange information in a dynamic, seamless manner almost anywhere and anytime. Knowledge extracted from social media has proven valuable in various applications. For example, real time analysis of Twitter data has been used to model earthquake warning detection systems [14], detect the spread of influenza-like-illness [15], predict the financial market movement [16,17], and identify potential product features for development of next generation products [18].

Despite the range of applications, design methodologies that leverage the power of social media data to mine information about products in the market are limited. Researchers in the design community have studied the importance of integrating lead users into the product development processes by recruiting customers for lead user studies [19] or mining product discussion blogs/reviews [20,21]. However, compared to social media driven models, existing lead user techniques may suffer from the following limitations:

(1) *Time and cost efficiencies*: gathering and understanding customer needs in a timely and efficient manner have been shown to be the single most important area of information necessary for product design and development [22]. In the

Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received September 15, 2014; final manuscript received February 9, 2015; published online xx xx, xxxx. Assoc. Editor: Wei Chen.

case of lead user needs identification, scouting and recruiting lead users can take several months [10]. In addition, lead user studies typically require participant compensation, hereby increasing costs and limiting the potential pool of participants. However, information in social media networks is readily available both to individuals seeking to post messages, and researchers seeking to acquire and store relevant product-related messages [18].

(2) *Homogeneity of lead user information*: the reliance on the physical presence of lead users during the lead user needs identification process potentially limits the heterogeneity of ideas. For example, design teams may have to travel and directly interact with lead users in a given geographic location, in order to acquire a heterogeneous perspective on existing product challenges [10]. For digital data such as web-blogs/product discussion forums, lead user preferences may be present [20]. However, the lack of geospatial information makes it difficult to verify the source and heterogeneity of the product-related information. Social media networks, on the other hand, enable users to provide geographically stamped identification [23] that can then be utilized by designers to discovery region-specific lead user preferences.

Recent works by Tuarob and Tucker have demonstrated the viability of using social media data to mine product features that customer express positive/negative sentiment toward [18]. The methodology presented in this work aims to discover product features that are not yet mainstream by identifying lead user messages within large scale social media networks. For example, messages that convey product ideas such as "U know with all the glass in the iPhone 4 they really should think about integrating a solar panel to recharge the battery." or "i wish i could use my iPhone as a universal remote control." are ubiquitous in social media. Hence, the ability to identify such product feature information from lead users in social media will help designers discovery emerging product features from individuals that are ahead of the technology market curve.

In this paper, the authors propose a data mining driven methodology that automatically identifies lead users of a particular product/product domain from a pool of social media users. In particular, the authors develop a set of algorithms that first identify latent features discussed in social media. The discovered latent features are then used to identify potential *product specific lead users* (lead users who have expertise in a particular product) and *global lead users* (lead users who have critical, innovative ideas that are applicable to all products within the product domain). This paper has the following main contributions:

(1) The authors adopt text mining techniques to extract product ground-truth features from product specification documents and user-discussed features from social media data.
(2) The authors propose a mathematical model to identify the latent features from the extracted ground-truth and user-discussed features.
(3) A probability-based mathematical model is developed to identify product specific and global lead users.
(4) The authors illustrate the efficacy of the proposed methodology using a case study of real world smart phone data and Twitter data.

The remainder of the paper is organized as follows: Sec. 2 discusses related literature. Section 3 discusses the proposed methodology used to address the challenges outlined above. Section 4 introduces the case study along with the experimental results and discussion. Section 5 concludes the paper.

2 Related Works

Literature on automatic identification of relevant product features is an emerging area of research, particularly due to advances in machine learning algorithms and computing infrastructure. The literature presented in this section includes research most closely related to the methodology presented in this paper.

2.1 Discovering Lead Users for Product Development. Lead user research in product design and development has primarily focused on discovering customers that provide innovative ideas that are ahead of market trends and preferences. Hippel et al. explored how lead users can be systematically discovered, and how lead user perceptions and preferences can be incorporated into industrial and customer marketing research analyses of emerging needs for new products, processes, and services [7–9,24]. Pia and Hölttä-Otto's research findings discovered that individuals with disabilities served as a valid source of lead users during the design of next generation cell phones [25]. Batallas et al. modeled and analyzed information flows within product development organizations [26]. The model leads to understanding and identifying *information leaders* in product development processes. Schreier et al. studied lead user participants and found that leaders have stronger, more innovative domain-specific ideas, compared to ordinary users. Moreover, they perceive new technologies as less complex and hence are in better positions to adopt them [27]. Vaughan et al. proposed a methodology to identify emphatic lead users from nonuser product design engineers through the use of simulated lead user experiences in order to mitigate the problems caused by the heterogeneity of culture, geographical location, and language among participants especially in the developing countries [28].

These works illustrate the benefits of lead users in providing innovative ideas during next generation product development efforts. However, acquiring such lead users can be time-consuming and costly [10] and may not reach out to all the potential lead users in the user space. The social media network model that is proposed in this work mitigates these challenges by enabling the automated discovery of lead users, in addition to the product features not yet realized by the existing customer pool.

2.2 Automatic Identification of Leaders. Literature in Computer Science and Information Retrieval has proposed methods to automatically identify *leaders* from pools of users in online communities. Zhao et al. proposed a machine learning based method to identify leaders in online cancer communities [29]. Their method is only applicable to specific cancer domains, as the learning process of the algorithm requires cancer specific domain knowledge. Song et al. proposed the *InfluenceRank* algorithm for identifying opinion leaders in Blogospheres [30]. Their algorithm utilizes networking connectivity among users which is not always available in some social media services. Tang et al. proposed the *UserRank* algorithm which combines link analysis and content analysis techniques to identify influential users in social network communities [31]. Multiple works have also been devoted to building automated systems to identify *leaders* or influential users in online communities such as in Refs. [32–34]. However, these existing techniques are not suitable for discovering lead user needs in large scale social networks due to: (1) most of the proposed algorithms in the literature require network structures among users which are not always available in social media services such as Twitter,[1] blogs, and product reviews and (2) the definition of *leaders* in most previous works pertains to how a user's opinion propagates (or *influences* other users) throughout the network, while a *lead user* in the product development sense is a user who experiences unknown needs. The differences in the definitions of a leader make previous algorithms motivate the development of the algorithms proposed in this work.

[1]Though one could infer the relationship among Twitter users by constructing communities based on the *Reply-To* connections, such connections are sparse and spurious. These are not taken into account in most network-based leader identification algorithms.

2.3 Product Feature Extraction. A critical step in customers' need acquisition is identifying product features relevant to next generation design and development, particularly relating to product feature mining in textual data.

Tucker and Kim proposed a machine learning based approach for mining product feature trends in the market from the time series of user preferences [35]. Their proposed model predicts future product trends and automatically classifies product features into three categories: *Obsolete*, *Nonstandard*, and *Standard* features. Other works by Tucker and Kim include mining publicly available customer review data for product features [36] and identifying relevant product features from a high dimensional feature set [37].

In extracting product features and opinions from textual data such as social media messages and product reviews, Popescu et al. proposed OPINE, an unsupervised system for extracting product features from user reviews [38]. Rai proposed a methodology for identifying key product attributes and their importance levels by mining online customer reviews [39]. Textual data are converted into a term document matrix and subsequently mined for product-related features. Ren and Papalambros proposed a crowd implicit feedback methodology for eliciting design preferences [40]. A black-box optimization approach is introduced to retrieve and update user preference models during the customer elicitation process. Stone and Choi proposed extracting customer preference from user-generated content based on machine learning classification. A support vector machine algorithm is employed to mine product attributes and their levels from online data [41]. Tuarob and Tucker proposed a topic modeling based feature extraction algorithm that takes a collection of social media messages related to a particular product as an input and extracts *strong*, *weak*, and *controversial* product features [18]. This approach works well with social media data; however, it cannot extract opinions associated with each extracted feature and does not scale well due to having to remodel topics every time new messages are added to the social media collection (i.e., the algorithm is nonupdateable). Huang et al. proposed a feature extraction algorithm as part the REVMINER[2] project, which mines restaurant reviews from the website[3] and summarizes the reviews to facilitate restaurant suggestion for travelers through a mobile app. The REVMINER feature extraction algorithm has two advantages over Tuarob and Tucker's algorithm in that: (1) it can continue to extract features from newly added data without having to run the whole process again and (2) it can extract opinions associated with each feature. Hence, the methodology in this work extends REVMINER's feature extraction algorithm to extract product features from noisy data under social media settings.

3 Methodology

The methodology begins by partitioning customer needs into *known* needs and *unknown* needs, along with whether those needs are *known* or *unknown* in the market (Fig. 1). Lead users represent quadrant two in Fig. 1, as their needs are known to them but unknown to the market. In this work, the authors outline an algorithm to extract lead user data from social media networks. In addition, product specification data are collected and serves as ground-truth validation of discovered lead user needs. The hypothesis that ~~lead users needs~~ exist in social media networks and are distinguishable from all other nonrelated product messages will be tested through the steps outlined in the following sections AQ2 of the methodology. Textual and temporal data, acquired from large scale social media data, serve as the data source to extract product-related features and mine for lead user needs.

3.1 Overview and Definitions. Figure 2 outlines the steps involved in the methodology. In this work, a product feature is defined as a noun phrase representing a property of a product. For

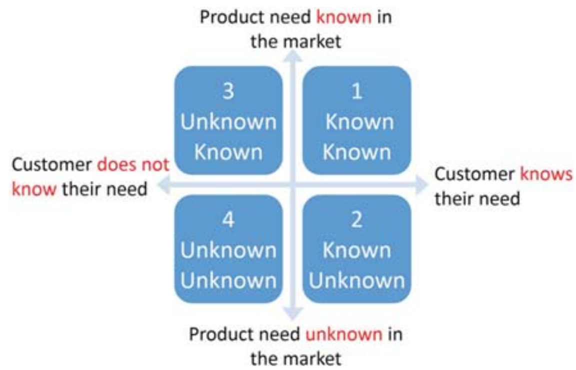

Fig. 1 Overview of the proposed methodology

example, features for smartphones include *screen*, *app*, *camera*, *battery life*, etc. Let $\mathbb{S}$ be the set of all products in a particular domain,[4] F be the set of all features, G be the set of all product specification documents, M be the set of all social media messages, and U be the set of all social media users. For a user $u \in U$, M_u is the set of social media messages composed by u. For $s \in \mathbb{S}$, G_s and M_s represent the set of specification documents and social media messages corresponding to product s, respectively. Similarly, $F(G_s)$ and $F(M_s)$ are the sets of product features extracted from G_s and M_s, respectively.

The first step in Fig. 2 is to collect and preprocess the product specification documents (G_s) and social media messages (M_s) for a product $s \in \mathbb{S}$. Then, the feature extractor algorithm extracts features from both sets of documents and produces a set of ground-truth product features $F(G_s)$ and a set of user-discussed product features $F(M_s)$. The ground-truth product features $F(G_s)$ represent product features existing in products on the market, clearly outlined in manufacturer specifications. The user-discussed product features $F(M_s)$, on the other hand, represent product features that are discussed by users in a social media network and may/may not exist in the current product offerings on the market. Therefore, $F(G_s)$ and $F(M_s)$ are used to identify the set of product specific latent features $F^*(s)$ and global latent feature $F^*(\mathbb{S})$. A *latent feature* is a product feature that has not been discovered or implemented in the market space. That is, such a feature is hidden from the market space. In this work, a latent feature is defined to be a product feature that is discussed in social media but does not yet exist in the market space. The last step of the methodology is to identify the lead users of each product s, and the global lead users across all the products in $\mathbb{S}$.

The primary challenge and fundamental contribution of this work is the automated classification of which latent features are relevant to a particular product need. For example, a ground-truth product feature set $F(G_s)$ for a cell phone product could be {bluetooth, ~~NFC~~, Lithium Ion Battery}, while the user-discussed product feature $F(M_s)$ could be {bluetooth, pillow, solar charging}. AQ3 For the product feature bluetooth, it exists in both the ground-truth product feature set $F(G_s)$ and user-discussed product feature $F(M_s)$ and would therefore be considered a standard feature already existing in the market. However, the product features *pillow*, *solar charging* expressed by users, are not part of the existing ground-truth product feature set $F(G_s)$. The fundamental challenge therefore becomes to develop an algorithm that automatically determines which product features represent latent product features relevant to the next generation phone product design and which product features are simply noise? The three main components (as shown in bold-gray *objective* boxes in Fig. 2) are presented to address this research objective.

[2] http://revminer.com/
[3] http://www.yelp.com/c/seattle/restaurants

[4] A product domain is a set of products that belong to the same category, e.g., *smartphone*, *automobile*, *laptop*, etc.

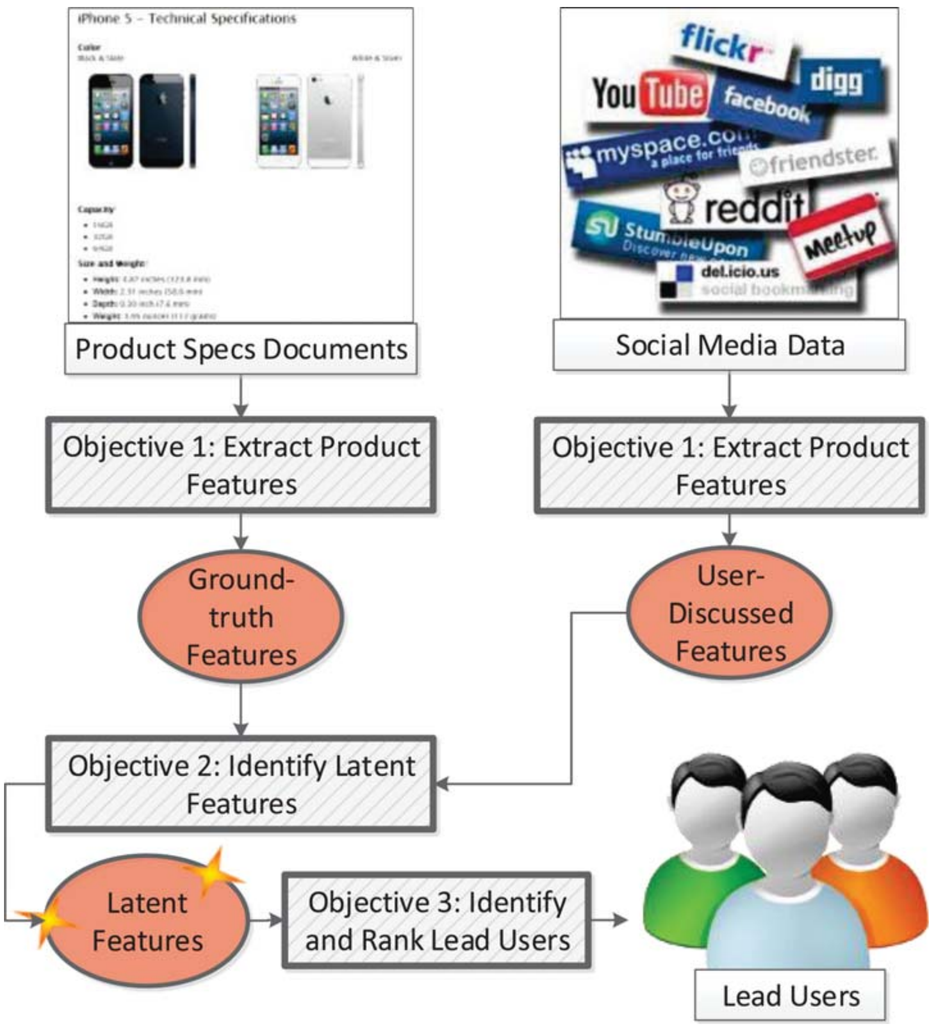

Fig. 2 Overview of the proposed methodology

3.2 Data Collection and Preprocessing

3.2.1 Collecting Product Specification Documents. A product specification document provides the actual nonbiased features of the product. These documents will be used to construct the ground-truth features for each chosen product and are primarily acquired from product technical specification manuals from the manufacturers.

3.2.2 Social Media Data Collection. Social media provides a means for people to interact, share, and exchange information and opinions in virtual communities and networks [42]. For generalization, the proposed methodology minimizes the assumption about functionalities of social media data, and only assumes that a unit of social media is a tuple of unstructured textual content, a user ~~ID~~, and a timestamp. Such a unit is referred to as a *message* throughout the paper. This minimal assumption would allow the proposed methodology to generalize across multiple heterogeneous pools of social media such as Twitter, Facebook, and Google+, as each of these social media platforms has this data structure.

3.3 Data Selection and Preprocessing. Social media messages, corresponding to each product domain, are retrieved by a query of the product's name (and variants) within the large stream of social media data. The technique developed by Thelwall et al. is employed to quantify the emotion in a message. The algorithm takes a short text as an input, and outputs two values, each of which ranges from 1 to 5 [43]. The first value represents the *positive* sentiment level, and the other represents the *negative* sentiment level. The reason for having the two sentiment scores instead of just one (with $-/+$ sign representing negative/positive sentiment) is because research findings have determined that positive and negative sentiment can coexist [44]. The positive and negative scores are then combined to produce an emotion strength score using the following equation:

$$\text{Emotion Strength (ES)} = \text{Positive Score} - \text{Negative Score} \quad (1)$$

Another reason for combining *Negative* and *Positive* scores is that messages with implicit sentiment (i.e., sarcasm) would be neutralized since such messages tend to have equally high volumes of both *Positive* and *Negative* scores, causing the *Emotion Strength* score to converge to 0 [45]. A message is then classified into one of the three categories based on the sign of the Emotion Strength score (i.e., positive (+ve), neutral (0ve), and negative (−ve)). The *Emotion Strength* scores will later be used to identify whether a particular message conveys a positive or negative attitude toward a particular product or product feature. The positive sentiment messages will then be used to approximate the demand of a particular product, as proposed in Ref. [18]. The approximated demand will be used in the computation of the ranking scores in order to find the global product lead users.

3.4 Objective 1: Product Feature Extraction From Textual Data. For each product $s \in \mathbb{S}$, the methodology extracts the ground-truth product features ($F(G_s)$) from the set of manufacturer-provided product specification documents (G_s)) that describe its actual features. Also extracted are the user-discussed features ($F(M_s)$) from the set of social media messages related to the product s (M_s). Since both G_s and M_s are collections of plain text documents, the same feature extraction algorithm is employed to mine product-related features.

Extracting product features from textual data proves to be one of the challenging extraction problems in the information retrieval (IR) literature. In this paper, a number of feature extraction algorithms proposed in Refs. [18,38,46–48] are considered. Out of these algorithms, the authors only have access to the core implementations of Refs. [18,47] and choose to extend the algorithm proposed by Huang et al. [47]. Though both feature extraction algorithms do not require domain knowledge about the products and are suitable for the focused task in this research, Huang et al.'s algorithm is extended due to its capability to process large, dynamic datasets with less computational time. The algorithm is also able to extract customers' opinions associated with each extracted feature.

Algorithm 1: The feature extraction algorithm from a collection of documents

---

Input: D: Set of free-text documents to extract product features.
Output: E: Set of extractions. Each $e \in E$ is a tuple of $\langle feature,\ opinion,\ frequency \rangle$, for example $e = \langle 'onscreen\ keyboard', 'fantastic', 5 \rangle$

```
preprocessing;
for d ∈ D do
    Clean d;
    POS tag d;
    Extract multi-word features;
end
initialization;
E = ∅;
T = ∅;
F = Seed Features;
while E can still grow do
    Learn templates from seed features;
    Add new template to T;
    foreach d ∈ D do
        foreach Sentence s ∈ d do
            e ← Extract potential feature-opinion pair using T;
            Add e to E;
        end
    end
    Update F;
end
E ← Clustering and normalizing features;
return E;
```

The original feature extraction algorithm proposed by Huang et al. was used to extract features of restaurants in the Seattle area from Yelp reviews [47]. The algorithm is enhanced in this work in order to handle noisy data more efficiently, such as that existing in social media. In particular, a data preprocessing step is added to clean residuals such as symbols, hyperlinks, usernames, and tags, and correct misspelled words. Such noise is ubiquitous in social media and could cause erroneous results [49]. This data cleaning process has shown to improve the performance of social media message classification by 6.3% on average [50]. The feature extraction algorithm used in this paper is outlined in Algorithm 1. The input is a collection of documents D that are textual in nature. Note that this can either be a collection of product specification documents (i.e., G_s) or a set of social media messages (i.e., M_s). The Stanford Part of Speech (POS) Tagger[5] is used to tag each word with an appropriate POS. This preprocessing step is required because a product feature is defined to be a noun phrase. The final step of the preprocessing phase extracts potential multiword features and stores them in a repository for subsequent mining. A multiword feature is a feature composed by two or more words such as *on-screen keyboard* and *Facebook notification*.

The core of the algorithm iteratively learns to identify features and generates a set of extractions (E) from the input collection of documents. Each extraction $e \in E$ is a tuple of $\langle$feature, *opinion*, *frequency*$\rangle$ such as $\langle 'onscreen\ keyboard', 'fantastic', 5 \rangle$, which infers that the *on-screen keyboard* feature of this specific product was mentioned as *fantastic 5* times within the product document corpus. The algorithm employs a bootstrapping method which is initialized with a small set of ground-truth features. The algorithm then repeatedly learns phrase templates surrounding the seed features and uses the templates to extract more features. As a simplified example to illustrate such a process, let *keyboard* be a ground-truth feature. When the algorithm comes across a textual message "... because of the new *keyboard* in iOS 7 ...," it memorizes the word pattern surrounding the word *keyboard*. If the algorithm ever comes across a similar sentence pattern again, e.g., "... because of the smart *text prediction* in iOS 8 ...," it would know that *text prediction* would also be a product feature. This process continues until the extraction set is no longer populated by new features. Additional details about the mechanics of the feature extraction algorithm can be found in Ref. [47].

Finally, the algorithm postprocesses the extractions by disambiguating and normalizing the features. The disambiguation process involves stemming the features using the Porter's stemming algorithm[6] and clustering them using the WordNet[7] SynSet. This postprocessing step groups the same features that may be written differently (e.g., `Screen`, `Monitor`, `Screens`, and `Monitors` would be grouped together).

Once the set of extractions E is generated, the $Feature(E)$ and $Opinion(E)$ are defined to be the sets of distinct features and distinct opinions, respectively. Hence given a collection of documents associated to a product s (either G_s or M_s), the feature extraction algorithm is able to extract a set of features related to the product (which is referred to as $F(G_s)$ or $F(M_s)$, respectively).

3.5 Objective 2: Identifying Latent Features. The proposed methodology defines a *latent feature* of a product domain as a feature that does not exist in any existing product within the domain. In other words, a latent feature is a feature that has not yet been implemented in any products in the market space. With such an assumption, one could automatically identify the set of latent features by subtracting the set of user-discussed features with the set of ground-truth features of all products. The authors define two types of latent features associated with lead users:

Product specific latent features ($F^*(s)$) are product features mined from lead users who may have innovative ideas pertaining to a specific product.

Global latent features ($F^*(\mathbb{S})$) are product features mined from lead users who have innovative product ideas that may be applicable across an entire product domain.

Product specific and global latent features will later be used to identify product specific and global lead users, respectively.

Mathematically, given a product domain $\mathbb{S}$, the set of product specific latent features of the product s, $F^*(s)$, and the set of global latent features $F^*(\mathbb{S})$ are defined as

$$F^*(\mathbb{S}) = \bigcup_{s \in \mathbb{S}} F(M_s) - \bigcup_{s \in \mathbb{S}} F(G_s) \tag{2}$$

$$F^*(s) = F(M_s) \cap F^*(\mathbb{S}) \tag{3}$$

In order to quantify the *meaningfulness* of each extracted latent feature (since some features could be just noise or remnants

[5] http://nlp.stanford.edu/downloads/tagger.shtml

[6] http://tartarus.org/martin/PorterStemmer/

[7] http://wordnet.princeton.edu/

caused by algorithmic flaws, such as "`http://`," "`i mean I`," etc.), the metric feature frequency-inverse product frequency (FF-IPF) is proposed, which is intuitively similar to the term frequency-inverse document frequency (TF-IDF) employed in the IR field [51]. In the IR field, TF-IDF is widely used for ranking words by their importance with respect to the documents in which it appears and the whole collection of documents. Another reason for transforming this problem into a traditional IR problem is that standard IR evaluation techniques and metrics could be applied [51–53]. TF-IDF has two components: the term frequency (TF) and the inverse document frequency (IDF). The TF is the frequency of a term appearing in a document. The IDF of a term measures how important the term is to the corpus and is computed based on the document frequency and the number of documents in which the term appears.

Similarly, a product can be textually described by a document (either technical manuals from manufacturers or social media messages). Based on this concept, a product is composed of a set of features, mined from Eqs. (2) and (3). A feature mining algorithm based on the TF-IDF metric would therefore quantify the importance of each feature of a product, relative to all features mined. If multiple products lack a certain feature that would satisfy a majority of lead users, then a high volume of discussion regarding the needs of such a feature would be expected. *Feature Frequency* quantifies this. On the other hand, if a latent feature is discussed sparsely amongst lead users, it is assumed that such a feature is rare, which can be quantified by the *Inverse Product Frequency*. Therefore, a meaningful latent feature (i.e., a latent feature that is simply not noise from the large scale social media data) is that which has a high frequency of being mentioned by a lead user and a low feature frequency across a wide range of products. Hence, a combined FF-IPF metric is used to quantify the meaningfulness of each latent feature. Mathematically, given a product collection $\mathbb{S}$ and a set of latent features $F^*(\mathbb{S})$, the FF, IPF, and FF-IPF of a latent feature $f \in F^*(\mathbb{S})$ are defined as

$$\mathrm{FF}(f, F^*) = 05 + 0.5 \times \frac{|\mathrm{Frequency}(f)|}{\sum_{f' \in F^*} |\mathrm{Frequency}(f')|} \quad (4)$$

$$\mathrm{IPF}(f, \mathbb{S}) = \log \frac{|\mathbb{S}|}{|\{s \in \mathbb{S} : f \in s\}|} \quad (5)$$

$$\mathrm{FF} - \mathrm{IPF}(f, F^*, \mathbb{S}) = \mathrm{FF}(f, F^*) \cdot \mathrm{IPF}(f, \mathbb{S}) \quad (6)$$

Note that the feature frequency score in Eq. (4) is constrained to the [0.5, 1] range to boost the score for features that occur less frequently, relative to features that are mentioned more frequently. This augment would consequently prevent the FF − IPF score to converge to zero (hence becoming nondiscriminative) for features that are less common [54]. The set of extracted latent features will be used to identify customers who possess and express innovative ideas, whom are referred to as *lead users*.

3.6 Objective 3: Identifying and Ranking Lead Users.

Berthon defines *lead users* as those who experience needs still unknown to the public and who also benefit greatly if they obtain a solution to these needs [55]. This section discusses how product specific and global lead users are identified and ranked from the heterogeneous pool of social media users. Recall that a *product specific lead user* is a customer who has expertise and is knowledgeable about a particular product; while a *global lead user* has critical and innovative ideas about all the products in a particular domain. For example, an iPhone-specific lead user who is familiar with multiple iPhone products may have a better sense of what innovative features could be incorporated into the next generation iPhone to specifically extend its capability to satisfy his/her needs (e.g., the lightning cable could be magnetized so it can snap into the charging port without much effort). On the other hand, a smartphone-global lead user may have tried or reviewed multiple smartphone products and is familiar with the boundaries of current smartphone inventions and is able to identify innovative features for the smartphone market in general (e.g., a smartphone could be used to replace credit cards when purchasing items). However, some of the innovative features that global lead users generate may not be compatible with particular smartphone products. A company would want product specific lead users' opinions to synthesize innovative features into their existing product lines, while global lead users' suggestions could give birth to a new groundbreaking product family that draws attention from customers whose needs are not met by current products in the market space.

3.6.1 Identifying Lead Users for a Particular Product. The proposed methodology automatically identifies lead users in a pool of social media users by detecting users who express innovative ideas about the products that they use or are familiar with. Specifically, given a user $u \in U$ and a product $s \in \mathbb{S}$, the methodology computes $P(u|s)$, the probability that the user u is a lead user of the product s. The probability is referred to as the product specific *iScore* (or the innovative score), which is a real number from [0,1] range and will be used later for ranking users. Top users with highest product specific *iScore*s are regarded as the *product specific lead users*.

Algorithm 2**:** Algorithm for identifying and ranking product specific lead users of a particular product s

Input: $s \in \mathbb{S}$: The product. U: The set of all users. $F(G_s)$:Ground-truth features. $F(M_s)$:User discussed features. $F^*(s)$: Latent features.
Output: Ranked list of users with respect to $P(u|s)$

```
initialization;
I = ∅;
foreach user u ∈ U do
    M_u ← The messages posted by u;
    Compute F(M_u) using Algorithm 1;
    iScore ← Compute P(u, s);
    Add ⟨u, iScore⟩ to I;
end
I ← Rank users in I by iScores;
return I
```

Algorithm 2 outlines the procedure of assigning a product specific *iScore* to a user, given a particular product s. $P(u|s)$ can be thought as the likelihood that the user u is a lead user for the product s and is defined as

$$P(u|s) = \sum_{f \in F(M_u)} P(u|f, s)(f|s) \quad (7)$$

where

$$P(u|f, s) = \begin{cases} 1; & f \in F^*(\mathbb{S}) \\ 0; & \text{Otherwise} \end{cases}, \quad P(f|s) = \frac{1}{|F(G_s) \cup F(M_s)|} \quad (8)$$

Equation (7) is directly expanded using the law of total probability, which sums over all the features expressed by the user u related to the product s, i.e., $F(M_u)$. $P(u|f, s)$ is the probability of the user u being the lead user, given a feature f, and is defined to be 1 if f is a latent feature, and 0 otherwise. Finally, $P(f|s)$ is the probability of a user expressing the feature f and can be computed directly from the pool of all features related to the product s. The current model assumes uniform distribution on the weights of the product features. That is, each product feature of the same product carries the same weight. Future work will explore possible weighting schemes for product features, so that users who mention critical features would be given higher probability $P(u|s)$ than those who mention only common features. Note that the value of $P(u|s)$ from Eq. (7) ranges between [0, 1].

3.6.2 Identifying Global Lead Users Within the Product Domain. In order to identify the global lead users across all the products in the product space $\mathbb{S}$, the global *iScore* (or $P(u)$) is computed for each user. Top users with highest global *iScores* are regarded as the global lead users of the product domain $\mathbb{S}$

$$P(u) = \sum_{s \in \mathbb{S}} P(u|s)(s) \quad (9)$$

Based on the law of total probability, $P(u)$ can be computed as the sum of proportional $P(u|s)$ across each product $s \in \mathbb{S}$. $P(s)$ is the probability of the product s being known and demanded by the market. Tuarob and Tucker found that the volume of the positive sentiment in social media corresponding to a particular product can be used to quantify the product demand which they found to directly correlate with the actual product sales [18]. In this work, the proposed methodology instantiates such findings and proposes to approximate $P(s)$ with the proportion of positive sentiment over all the products in the same domain, i.e.,

$$P(s) = \frac{|\text{Positive}(s)|}{\sum_{s' \in \mathbb{S}} |\text{Positive}(s')|} \quad (10)$$

Positive(s) is the set of positive messages associated with the product s. Note that the value of $P(s)$ from Eq. (9) ranges between [0, 1].

4 Case Study, Results, and Discussion

This section introduces a case study used to verify the proposed methodology and discusses the results.

4.1 Case Study. A case study of 27 smartphone products is presented that uses social media data (Twitter data) to mine relevant product design information. Data pertaining to product specifications from the smartphone domain are then used to validate the proposed methodology. The selected smartphone models include *BlackBerry Bold 9900*, *Dell Venue Pro*, *HP Veer*, *HTC ThunderBolt*, *iPhone 3G*, *iPhone 3GS*, *iPhone 4*, *iPhone 4S*, *iPhone 5*, *iPhone 5C*, *iPhone 5S*, *Kyocera Echo*, *LG Cosmos Touch*, *LG Enlighten*, *Motorola Droid RAZR*, *Motorola DROID X2*, *Nokia E7*, *Nokia N9*, *Samsung Dart*, *Samsung Exhibit 4G*, *Samsung Galaxy Nexus*, *Samsung Galaxy S 4G*, *Samsung Galaxy S II*, *Samsung Galaxy Tab*, *Samsung Infuse 4G*, *Sony Ericsson Xperia Play*, and *T-Mobile G2x*.

Smartphones are used as a case study in this paper because of the large volume of discussion about this product domain in social media. Previous work also illustrated that social media data (i.e., Twitter) contain crucial information about product features of other more mundane products such as automobiles [56]. The proposed algorithms may not work well for products which are not prevalent in social media as the corresponding sets of social media messages may be too small to extract useful knowledge from.

4.1.1 Smartphone Specification Data. The ground-truth specifications of each smartphone model are collected from product specification manual provided by the manufacturer (as a ~~PDF~~ document) or publicly available online. These documents are downloaded by the authors. Only textual information is extracted from each product specification document since the feature extraction algorithm employed in this research works primarily with textual data.

4.1.2 Product-Related Twitter Data. Twitter[8] is a microblog service that allows its users to send and read text messages of up to 140 characters, known as *tweets*. The Twitter dataset used in this research was collected randomly using the provided Twitter API and comprises 2,117,415,962 (2.1×10^9) tweets in the United States during the period of 31 months, from March 2011 to September 2013.

Tweets related to a product are collected by detecting the presence of the product name (and variants) and preprocessed by cleaning and mapping sentiment level as discussed in Sec. 3.3. Table 1 lists the number of tweets, percentage positive sentiment, and number of unique Twitter users of each chosen smartphone model. The percentage positive sentiment of a product s is calculated by $(|\text{Positive}(s)|/|\text{AllTweets}(s)|) \times 100\%$, where *Positive*$(s)$ is the number of positive tweets related to the product s.

Figure 3 displays the monthly Twitter discussion share of each chosen smartphone model throughout the 31 month period of data collection. Note that, since some smartphone models (i.e., the *iPhones*) have enormous discussion shares compared to other cell phone products, the methodology normalizes the social media messages accordingly.

4.2 Objective 1: Product Feature Extraction From Textual Data. Given a product $s \in \mathbb{S}$, the feature extraction algorithm (see Algorithm 1) is applied to the product specification documents (G_s) in order to obtain the ground-truth features $(F(G_s))$ and to the tweets related to the product (M_s) in order to extract features discussed by the Twitter users $(F(M_s))$. Table 2 enumerates the number of extracted ground-truth features, number of user-discussed features, and number of product specific latent features. Recall that a product specific latent feature of a product s is a feature mentioned in the set of social media messages related to s and does not appear in ground-truth features of any products in the product space $\mathbb{S}$.

4.3 Objective 2: Identifying Latent Features. A set of 25,816 global latent features $(F^*(\mathbb{S}))$ are extracted from the smartphone related social media data. A FF-IPF score is calculated for each latent feature. Figure 4 plots the distribution of the FF-IPF scores using a histogram, with an average-moving trend line. The distribution is heavily skewed to the right, suggesting an exponential growth. This would mean that a majority of the extracted latent features are meaningful (i.e., not noisy and erroneous features). Latent features with FF-IPF scores lower than 1.1 are treated as noise and eliminated, leaving with a set of 22,285 global latent feature for further processing.

Table 1 Selected smartphone models, their associated number of tweets, proportion of positive sentiment tweets (in %), and number of unique users who posted these tweets

Model	NumTweets	% Positive	NumUsers
BlackBerry Bold 9900	308	36.04	252
Dell Venue Pro	96	46.88	64
HP Veer	143	31.47	110
HTC ThunderBolt	1157	30.68	851
iPhone 3G	2154	25.63	1874
iPhone 3GS	3803	28.06	3119
iPhone 4	68860	28.92	43957
iPhone 4S	63500	29.53	39145
iPhone 5	211311	28.66	124461
iPhone 5C	5533	24.62	4475
iPhone 5S	15808	26.45	12417
Kyocera Echo	52	26.92	42
LG Cosmos Touch	23	39.13	20
LG Enlighten	18	16.67	17
Motorola Droid RAZR	2535	32.54	1981
Motorola DROID X2	471	26.75	378
Nokia E7	26	30.77	18
Nokia N9	208	34.13	153
Samsung Dart	29	20.69	28
Samsung Exhibit 4G	23	39.13	22
Samsung Galaxy Nexus	5218	31.07	2988
Samsung Galaxy S 4G	188	31.91	152
Samsung Galaxy S II	4599	31.12	3517
Samsung Galaxy Tab	3989	30.96	2578
Samsung Infuse 4G	284	34.15	215
Sony Ericsson Xperia Play	481	26.20	325
T-Mobile G2x	83	32.53	69

[8]https://twitter.com/

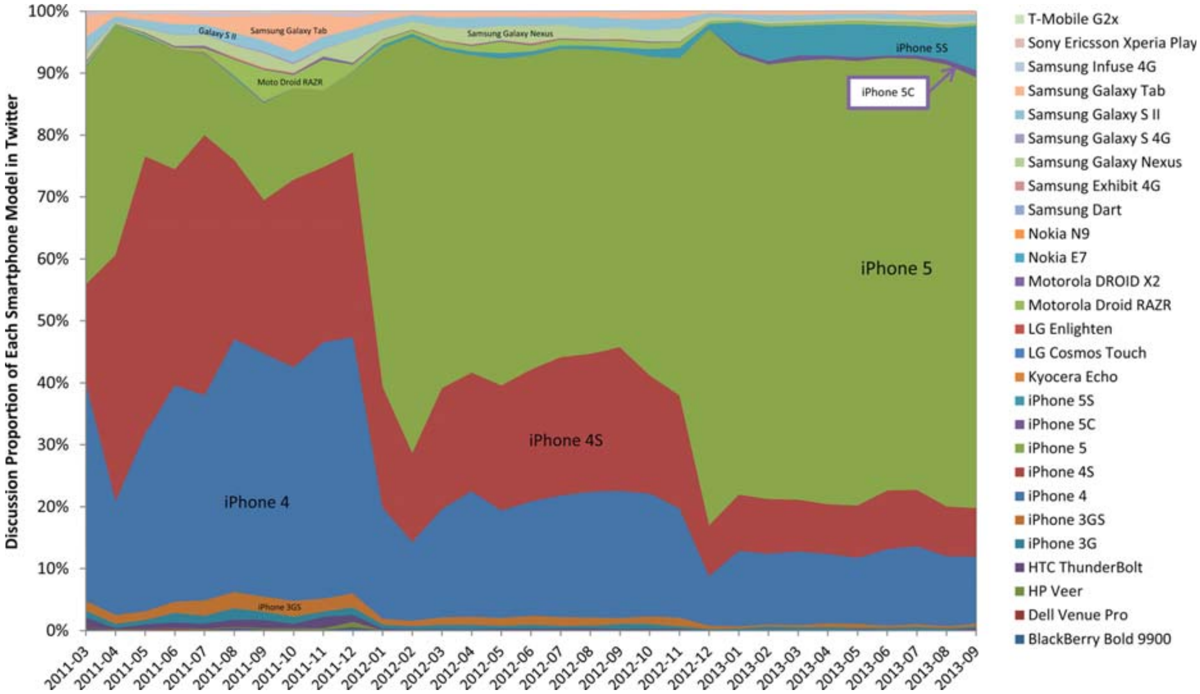

Fig. 3 Monthly distribution of Twitter discussion of each smartphone model across the 31 month period of data collection

Table 3 lists the top five extracted global latent features with highest FF-IPF scores, along with the tweets that provide contextual information about such latent features. These top five latent features reflect the actual customers' needs that have not been satisfied. These innovative opinions (as interpreted from the sample tweet associated with each latent feature) could be critical when designing next generation products. For example, customers

Table 2 Numbers of extracted ground-truth (base) features, user-discussed (user) features, and product specific latent features of each smartphone model

Model	# Base features	# User features	# Latent features
BlackBerry Bold 9900	1126	126	101
Dell Venue Pro	497	50	36
HP Veer	1206	76	56
HTC ThunderBolt	627	335	281
iPhone 3G	1330	532	420
iPhone 3GS	891	775	652
iPhone 4	995	6057	5720
iPhone 4S	963	5922	5582
iPhone 5	1020	13,493	13,050
iPhone 5C	895	833	717
iPhone 5S	973	1962	1740
Kyocera Echo	895	22	16
LG Cosmos Touch	769	11	6
LG Enlighten	1084	5	1
Motorola Droid RAZR	582	593	496
Motorola DROID X2	504	162	138
Nokia E7	749	14	10
Nokia N9	745	83	62
Samsung Dart	1178	10	6
Samsung Exhibit 4G	1331	10	7
Samsung Galaxy Nexus	456	1147	1017
Samsung Galaxy S 4G	1322	62	37
Samsung Galaxy S II	1319	801	662
Samsung Galaxy Tab	771	884	762
Samsung Infuse 4G	1121	85	60
Sony Ericsson Xperia Play	726	132	102
T-Mobile G2x	945	39	23

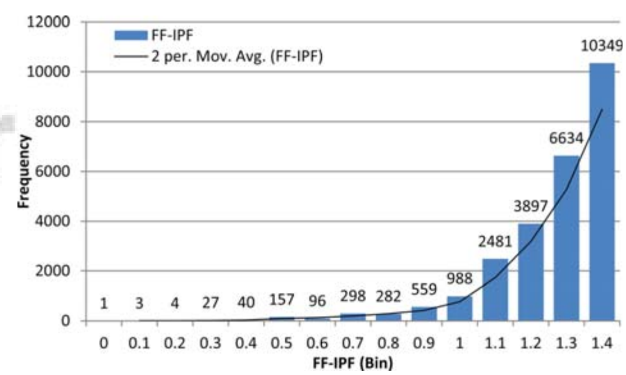

Fig. 4 Histogram showing the distribution of the FF-IPF scores of 25,816 total extracted global latent features

express needs for the *waterproof* feature for their *iPhones*; some users believe that a *solar panel* could be embedded underneath the *iPhone* screen so that the phone could charge itself when exposed to sunlight; etc. Note that the latent feature *hybrid* in the given example could be interpreted as either energy-source related or physical-feature related. This problem arises when a feature term is used to refer to more than one distinct features and paves the path to future works on semantic disambiguation of feature representation.

Figure 5 illustrates the proportion of tweets that mention the *waterproof* feature. From the plot, since the Twitter data were collected after March 2011, it is possible that the *waterproof* feature could have first been mentioned earlier than March 2011. However, the first model of dedicated waterproof smartphones (i.e., *Sony Xperia Z*) was not launched until early 2013.[9] Hence, the ability to identify critical latent features that would become manufacturable in the future could give designers advantages against the competitors in the market.

[9] http://www.tntmagazine.com/news/world/sony-announce-the-worlds-first-waterproof-phone-the-xperia-z

Table 3 Top five latent features across the chosen smartphone models, FF-IPF scores, and example tweets that related to the latent features

Latent feature	FF-IPF	Example
Waterproof	1.3087	I hope Apple incorporates some of that new ***waterproof*** technology in the iPhone 5 iPhone 5 better be ***waterproof***, shockproof, scratchproof, thisproof, thatproof, and all the rest of the proofs for $800
Solar panel	1.3061	... and what else would make the iPhone 5 even better, built in ***solar power charging***! U know with all the glass in the iPhone 4 they really should think about integrating a ***solar panel*** to recharge the battery.
Hybrid	1.3027	I wish there was an #android phone out there that was a ***hybrid of the best features on the droid razr maxx and the galaxy nexus***. I need a ***hybrid-iPhone4s*** so the battery can hold on all day when I'm at #vmworld. Steve, are you listening?:).
Tooth pick	1.3023	I hope iPhone 5 borrows from Swiss Army and finally adds a ***removable tooth pick***.
iHome	1.3021	My life would be 827492916 times better if my ***iHome took my iPhone 5*** First world problem: mad because my ***iPhone 5 is not compatible with this iHome*** dock in the hotel room.

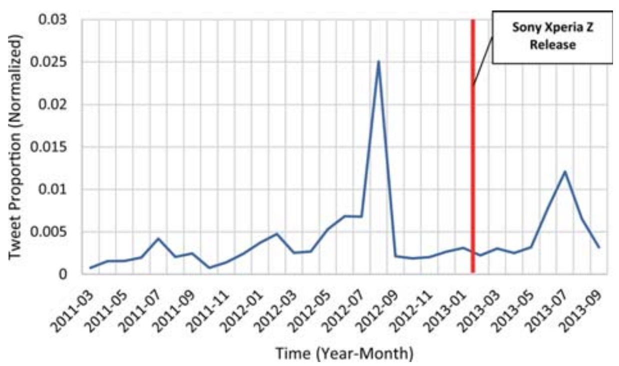

Fig. 5 Proportion of smartphone tweets which discuss the *waterproof* feature

4.4 Objective 3: Identifying and Ranking Lead Users. Once a set of latent features ($F^*(\mathbb{S})$) is identified, the product specific and global *iScore*s can be computed for each user in order to identify both product specific and global lead users.

For each product s, $P(u|s)$ is computed for each of the users in the pool of 198,974 Twitter users who tweet about their smartphone products according to Eq. (7). Then, $P(u)$ is computed according to Eq. (9). Table 4 lists some Twitter comments of the top lead user of each sample five smartphone models (i.e., *Samsung Galaxy Nexus*, *HTC ThunderBolt*, *iPhone 5*, *Sony Ericsson Xperia Play*, and *Kyocera Echo*). These tweets contain innovative ideas for improving these products. For example, a lead user suggests that the *Siri* functionality in the *iPhone 5* should be able to do more than just talk (he might be suggesting that the *iPhone 5* could connect to external hardware to enable *Siri* to perform physical interactions). Furthermore, one lead user of the *Sony Ericsson Xperia Play*, a smartphone that emphasizes on the gaming functionality, suggests to incorporate the ability to use the *Playstation 3* controllers with the phone.

These product specific lead users experience needs to improve the products during product usage. Identifying such product specific lead users would enable designers to seek solutions and innovative ideas for their next generation products across a wide range of users in a timely and efficient manner.

Oftentimes a lead user can be critical about product features across multiple products (not just his/her own products). Identifying these global lead users could bring out experts that could give better critical product development ideas. For this, all the users are ranked based on the $P(u)$ scores. Table 5 lists Twitter messages posted by the top global lead user with highest global *iScore*s that infer innovative ideas about smartphone features.

Table 4 Sample tweets from the top lead user of each sample five smartphone models. These tweets suggest product innovative improvement for each corresponding product.

Model	Product iScore	Sample Twitter message
Samsung Galaxy Nexus	0.0496	I wish there was an #android phone out there that was a ***hybrid*** of the best features on the droid razr maxx and the galaxy nexus.
HTC ThunderBolt	0.0308	HTC Thunderbolt #fail: Connect phone to PC to access drivers on included ***SD card*** ... but need drivers installed to access SD card from PC
iPhone 5	0.0174	but unless ***Siri*** can do more that just talk ...I'm not sold! #iPhone5
Sony Ericsson Xperia Play	0.0085	Hmm.. Playing games supporting Xperia Play controls. Wish I could use ***PS3 controller*** .. Makes me want an LTE Xperia Play with Tegra3..
Kyocera Echo	0.0077	Kyocera Echo needs to develop its own ***apps***.

Table 5 Sample tweets from the top five global lead users of the smartphone domain. These tweets suggest product innovation.

Global iScore	Sample Twitter message
0.0127	I wish there were a tweak for the iPhone 4S that would *indicate "4G" instead of just 3G* when I'm connected with a HSDPA + connection.
0.0126	If you trust my instinct, the iPhone 5S will come in *multiple colors and two display sizes*
0.0113	Very exciting Siri on the iPhone 4S *activates when you "raise it to your ear"* that'd b awesome.
0.0107	I wish i could use my iPhone as a *universal remote control*.
0.0105	Since iPhone already does fingerprint, Sumsung should *scan eyes*.

5 Conclusions and Future Works

This paper presents a data mining driven methodology to identify innovative customers, or *lead users*, from a heterogeneous pool of social media users. The methodology comprises of three main steps. First, product ground-truth features are extracted from the product specification documents, and the user-discussed features are extracted from social media data. Second, latent features (unrealized features) are extracted from the ground-truth and user-discussed features across all the products in the product space. Third, the product specific and global innovative scores (*iScores*) are computed for each user in the user space. Top product specific users are then regarded as the lead users of such a product. Also, users with top global *iScore*s are regarded as the global lead users. A case study of real-world 27 smartphone models with 31 month's worth of Twitter data is presented. The results and selected examples not only establish social media as a potential source for knowledge beneficial to product development and design, but also demonstrate that it is possible to build an automated system that identifies potential lead users from the pool of social media users along with potential latent features that they generate. This knowledge could be useful for development of next generation products. Future works could strengthen the evaluation process by involving user studies and verify the generalizability of the proposed methods by examining diverse case studies of different product domains and social media services, along with investigating the use of geographical information to mine lead users' preferences in different regions. Machine learning based techniques that allow multiple machines to learn different aspects of social media data such as Refs. [50,57–60] could be applied to enhance the performance of the feature extraction algorithm.

References

[1] Selden, L., and MacMillan, I. C., 2006, "Manage Customer-Centric Innovation—Systematically," Harv. Bus. Rev., **84**(9), pp. 149–150.

[2] Shah, S., 2000, "Sources and Patterns of Innovation in a Consumer Products Field: Innovations in Sporting Equipment," *Sloan School of Management*, Massachusetts Institute of Technology, Cambridge, MA, WP-4105.

[3] Tietz, R., Morrison, P. D., Luthje, C., and Herstatt, C., 2005, "The Process of User-Innovation: A Case Study in a Consumer Goods Setting," Int. J. Prod. Dev., **2**(4), pp. 321–338.

[4] Luthje, C., 2004, "Characteristics of Innovating Users in a Consumer Goods Field: An Empirical Study of Sport-Related Product Consumers," Technovation, **24**(9), pp. 683–695.

[5] Franke, N., Von Hippel, E., and Schreier, M., 2006, "Finding Commercially Attractive User Innovations: A Test of Lead-User Theory," J. Prod. Innovation Manage., **23**(4), pp. 301–315.

[6] Baldwin, C., and Von Hippel, E., 2010, "Modeling a Paradigm Shift: From Producer Innovation to User and Open Collaborative Innovation," Harvard Business School Finance Working Paper, Paper No. 10-038, pp. 4764–4809.

[7] Von Hippel, E., 1986, "Lead Users: A Source of Novel Product Concepts," Manage. Sci., **32**(7), pp. 791–805.

AQ8 [8] von Hippel, E., Ogawa, S., and de Jong Jeroen, P., 2011, "The Age of the Consumer–Innovator," MIT Sloan Manage. Rev., **53**(1), pp. 27–35.

[9] Herstatt, C., and von Hippel, E., 1992, "From Experience: Developing New Product Concepts Via the Lead User Method: A Case Study in a "Low-Tech" Field," J. Prod. Innovation Manage., **9**(3), pp. 213–221.

[10] von Hippel, E., Thomke, S., and Sonnack, M., 1999, "Creating Breakthroughs at 3M," Harv. Bus. Rev., **77**(5), pp. 47–57.

[11] Lilien, G. L., Morrison, P. D., Searls, K., Sonnack, M., and Hippel, E. V., 2002, "Performance Assessment of the Lead User Idea-Generation Process for New Product Development," Manage. Sci., **48**(8), pp. 1042–1059.

[12] Wu, X., Zhu, X., Wu, G.-Q., and Ding, W., 2014, "Data Mining With Big Data," IEEE Trans. Knowl. Data Eng., **26**(1), pp. 97–107.

[13] Corporation, I., 2013, "What is Big Data?—Bringing Big Data to the Enterprise," Accessed Aug. 16, 2013, http://www-01.ibm.com/software/ph/data/bigdata/

[14] Sakaki, T., Okazaki, M., and Matsuo, Y., 2010, "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," Proceedings of the 19th International Conference on World Wide Web, WWW'10, ACM, pp. 851–860.

[15] Collier, N., and Doan, S., 2012, "Syndromic Classification of Twitter Messages," *Electronic Healthcare* (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), P. Kostkova, M. Szomszor, and D. Fowler, eds., Vol. 91, Springer, Berlin, Germany, pp. 186–195.

[16] Bollen, J., Mao, H., and Zeng, X., 2011, "Twitter Mood Predicts the Stock Market," J. Comput. Sci., **2**(1), pp. 1–8.

AQ4 [17] Zhang, X., Fuehres, H., and Gloor, P., 2012, "Predicting Asset Value Through Twitter Buzz," *Advances in Collective Intelligence 2011*, Springer, ■, pp. 23–34.

[18] Tuarob, S., and Tucker, C. S., 2013, "Fad or Here to Stay: Predicting Product Market Adoption and Longevity Using Large Scale, Social Media Data," Proceedings of ASME 2013 International Design Engineering Technical Conference, Computers and Information in Engineering Conference (IDETC/CIE2013), ■. AQ5

[19] Lin, J., and Seepersad, C. C., 2007, "Empathic Lead Users: The Effects of Extraordinary User Experiences on Customer Needs Analysis and Product Redesign," ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. 289–296.

[20] Droge, C., Stanko, M. A., and Pollitte, W. A., 2010, "Lead Users and Early Adopters on the Web: The Role of New Technology Product Blogs," J. Prod. Innovation Manage., **27**(1), pp. 66–82.

[21] Bilgram, V., Brem, A., and Voigt, K.-I., 2008, "User-Centric Innovations in New Product Development-Systematic Identification of Lead Users Harnessing Interactive and Collaborative Online-Tools," Int. J. Innovation Manage., **12**(03), pp. 419–458.

[22] Ogawa, S., and Piller, F. T., 2006, "Reducing the Risks of New Product Development," MIT Sloan Manage. Rev., **47**(2), pp. 65–71.

[23] Bodnar, T., Tucker, C., Hopkinson, K., and Bilen, S., 2014, "Increasing the Veracity of Event Detection on Social Media Networks Through User Trust Modeling," Proceedings of the 2014 IEEE International Conference on Big Data, Institute of Electrical and Electronics Engineers, pp. 289–296.

[24] Von Hippel, E., 1978, "Successful Industrial Products From Customer Ideas," J. Mark., **42**(1), pp. 39–49.

[25] Hannukainen, P., and Hölttä-Otto, K., 2006, "Identifying Customer Needs: Disabled Persons as Lead Users," ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. 243–251.

[26] Batallas, D., and Yassine, A., 2006, "Information Leaders in Product Development Organizational Networks: Social Network Analysis of the Design Structure Matrix," IEEE Trans. Eng. Manage., **53**(4), pp. 570–582.

[27] Schreier, M., Oberhauser, S., and Prügl, R., 2007, "Lead Users and the Adoption and Diffusion of New Products: Insights From Two Extreme Sports Communities," Mark. Lett., **18**(1–2), pp. 15–30.

[28] Vaughan, M. R., Seepersad, C. C., and Crawford, R. H., 2014, "Creation of Empathic Lead Users From Non-Users Via Simulated Lead User Experiences," Proceedings of the ASME 2014 International Design Engineering Technical Conference, Computers and Information in Engineering Conference (IDETC/CIE2014), ■.

[29] Zhao, K., Qiu, B., Caragea, C., Wu, D., Mitra, P., Yen, J., Greer, G. E., and Portier, K., 2011, "Identifying Leaders in an Online Cancer Survivor Community," Proceedings of the 21st Annual Workshop on Information Technologies and Systems (WITS'11), pp. 115–120.

[30] Song, X., Chi, Y., Hino, K., and Tseng, B., 2007, "Identifying Opinion Leaders in the Blogosphere," Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM'07, ACM, pp. 971–974.

[31] Tang, X., and Yang, C., 2010, "Identifying Influential Users in an Online Healthcare Social Network," 2010 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 43–48.

[32] Li, Y.-M., Lin, C.-H., and Lai, C.-Y., 2010, "Identifying Influential Reviewers for Word-of-Mouth Marketing," Electron. Commer. Res. Appl., **9**(4), pp. 294–304.

[33] Trusov, M., Bodapati, A. V., and Bucklin, R. E., 2010, "Determining Influential Users in Internet Social Networks," J. Mark. Res., **47**(4), pp. 643–658.

[34] Aral, S., and Walker, D., 2012, "Identifying Influential and Susceptible Members of Social Networks," Science, **337**(6092), pp. 337–341.

[35] Tucker, C., and Kim, H., 2011, "Trend Mining for Predictive Product Design," ASME J. Mech. Des., **133**(11), p. 111008.

[36] Tucker, C. S., and Kim, H. M., 2009, "Data-Driven Decision Tree Classification for Product Portfolio Design Optimization," ASME J. Comput. Inf. Sci. Eng., **9**(4), p. 041004.

[37] Tucker, C., and Kim, H., 2011, "Predicting Emerging Product Design Trend by Mining Publicly Available Customer Review Data," Proceedings of the 18th International Conference on Engineering Design (ICED11), Vol. 6, pp. 43–52.

[38] Popescu, A.-M., and Etzioni, O., 2005, "Extracting Product Features and Opinions From Reviews," Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT'05, Association for Computational Linguistics, pp. 339–346.

[39] Rai, R., 2012, "Identifying Key Product Attributes and Their Importance Levels From Online Customer Reviews," ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. 533–540.

[40] Ren, Y., and Papalambros, P. Y., 2012, "On Design Preference Elicitation With Crowd Implicit Feedback," ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. 541–551.

[41] Stone, T., and Choi, S.-K., 2013, "Extracting Consumer Preference From User-Generated Content Sources Using Classification," ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, ■.

[42] Ahlqvist, T., and Teknillinen Tutkimuskeskus, V., 2008, *Social Media Roadmaps: Exploring the Futures Triggered by Social Media* (VTT Tiedotteita—Research Notes), No. 2454, VTT, ■.

[43] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A., 2010, "Sentiment in Short Strength Detection Informal Text," J. Am. Soc. Inf. Sci. Technol., **61**(12), pp. 2544–2558.

[44] Fox, E., 2008, *Emotion Science: Cognitive and Neuroscientific Approaches to Understanding Human Emotions*, Palgrave Macmillan, ■.

[45] Thelwall, M., 2013, "Heart and Soul: Sentiment Strength Detection in the Social Web With Sentistrength," Proceedings of the CyberEmotions, pp. 1–14.

[46] Tuarob, S., and Tucker, C. S., 2014, "Discovering Next Generation Product Innovations by Identifying Lead User Preferences Expressed Through Large Scale Social Media Data," Proceedings of ASME International Design Engineering Technical Conferences & Computers and Information in Engineering Conference 2014, ASME, ■.

[47] Huang, J., Etzioni, O., Zettlemoyer, L., Clark, K., and Lee, C., 2012, "RevMiner: An Extractive Interface for Navigating Reviews on a Smartphone," Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology, UIST'12, ACM, pp. 3–12.

[48] Hu, M., and Liu, B., 2004, "Mining Opinion Features in Customer Reviews," Proceedings of the 19th National Conference on Artificial Intelligence, AAAI'04, AAAI Press, pp. 755–760.

[49] Yin, P., Ram, N., Lee, W.-C., Tucker, C., Khandelwal, S., and Salathé, M., 2014, "Two Sides of a Coin: Separating Personal Communication and Public Dissemination Accounts in Twitter," *Advances in Knowledge Discovery and Data Mining*, Springer, ■, pp. 163–175.

AQ9 [50] Tuarob, S., Tucker, C. S., Salathe, M., and Ram, N., 2014, "An Ensemble Heterogeneous Classification Methodology for Discovering Health-Related Knowledge in Social Media Messages," J. Biomed. Inf., **49**, pp. 255–268.

[51] Manning, C. D., Raghavan, P., and Schütze, H., 2008, *Introduction to Information Retrieval*, Cambridge University Press, New York.

[52] Tuarob, S., Pouchard, L. C., and Giles, C. L., 2013, "Automatic Tag Recommendation for Metadata Annotation Using Probabilistic Topic Modeling," Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL'13, ACM, pp. 239–248.

[53] Tuarob, S., Bhatia, S., Mitra, P., and Giles, C. L., 2013, "Automatic Detection of Pseudocodes in Scholarly Documents Using Machine Learning," 2013 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 738–742. AQ6

[54] Sehgal, A., and Iowa Computer Science, T. U., 2007, *Profiling Topics on the Web for Knowledge Discovery*, University of Iowa, Iowa City, IA.

[55] Berthon, P. R., Pitt, L. F., McCarthy, I., and Kates, S. M., 2007, "When Customers Get Clever: Managerial Approaches to Dealing With Creative Consumers," Bus. Horiz., **50**(1), pp. 39–47.

[56] Tuarob, S., and Tucker, C. S., "Quantifying Product Favorability and Extracting Notable Product Features Using Large Scale Social Media Data," ASME J. Comput. Inf. Sci. Eng. (in press). AQ7

[57] Tuarob, S., Tucker, C. S., Salathe, M., and Ram, N., 2013, "Discovering Health-Related Knowledge in Social Media Using Ensembles of Heterogeneous Features," Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, CIKM'13, ACM, pp. 1685–1690.

~~[58] Tuarob, S., Bhatia, S., Mitra, P., and Giles, C., 2013, "Automatic Detection of Pseudocodes in Scholarly Documents Using Machine Learning," 2013 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 738–742.~~

[59] Bhatia, S., Tuarob, S., Mitra, P., and Giles, C. L., 2011, "An Algorithm Search Engine for Software Developers," Proceedings of the 3rd International Workshop on Search-Driven Development: Users, Infrastructure, Tools, and Evaluation, SUITE'11, ACM, pp. 13–16.

[60] Tuarob, S., Mitra, P., and Giles, C. L., 2012, "Taxonomy-Based Query-Dependent Schemes for Profile Similarity Measurement," Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search, JIWES'12, ACM, pp. 8:1–8:6.