Quantifying Product Favorability and Extracting Notable Product Features Using Large Scale Social Media Data

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 Engineering,
Computer Science and Engineering,
The Pennsylvania State University,
University Park, PA 16802
e-mail: ctucker4@psu.edu

Some of the challenges that designers face in getting broad external input from customers during and after product launch include geographic limitations and the need for physical interaction with the design artifact(s). Having to conduct such user-based studies would require huge amounts of time and financial resources. In the past decade, social media has emerged as an increasingly important medium of communication and information sharing. Being able to mine and harness product-relevant knowledge within such a massive, readily accessible collection of data would give designers an alternative way to learn customers' preferences in a timely and cost-effective manner. In this paper, we propose a data mining driven methodology that identifies product features and associated customer opinions favorably received in the market space which can then be integrated into the design of next generation products. Two unique product domains (smartphones and automobiles) are investigated to validate the proposed methodology and establish social media data as a viable source of large scale, heterogeneous data relevant to next generation product design and development. We demonstrate in our case studies that incorporating suggested features into next generation products can result in favorable sentiment from social media users. [DOI: 10.1115/1.4029562]

1 Introduction

A *product feature* is defined as an attribute of a product that is of interest to customers [1]. Product features that are well aligned with customer needs amplify their popularity in the market space and result in subsequent successes of future product iterations. On the other hand, products that are not well aligned with customers' needs may result in negative word of mouth feedback that may influence future potential purchasing decisions and subsequently result in discontinuation of the product lines [2]. Hence, designing product features relevant to market trends is a crucial step in the product design and development process. However, the advent of global competitive markets makes modeling trends difficult. Recent studies have shown that involving customers in product development process is more effective than perceiving them as the end of the product chains [3–5]. However, having customers' direct input has typically required them to either be physically present with the design teams during the prototype evaluation process or prototypes be sent out to their locations [6], thereby severely limiting the size, heterogeneity, and quality of customers that can evaluate the potential success of a design artifact. As a result, a substantial number of products that are purchased by customers each year are returned, resulting in wasted design efforts, wasted natural resources, and a decrease in long term customer satisfaction.

Society generates more than 2.5 quintillion (10^{18}) bytes of data each day [7,8]. 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 [9]. 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 [10], identify medical and emergency needs during recovery from natural disasters (such as the Haiti Earthquake) [11], detect the spread of influenza-like-illness [12], predict the financial market movement [13], and recommend products [14].

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 proposed using web-blogs or product review sites to mine product information due to the predefined categories of opinions and completeness of the information [15]; however, such website-based information may suffer from the following limitations:

(1) Immediacy: Website-based content, especially product review blogs, usually takes longer time for prepublishing processes including verifying content and proofreading, hereby possibly making the information out-of-date by the time it is available to the public [16]. The problem is further magnified in the case of time-sensitive products such as mobile apps and software packages where next releases or "patches" can take hours for development. Social media, on the other hand, promotes timeliness which allows its users to express their opinions which are immediately available.
(2) Reach: The amount of data available to designers may be limited due to designers' predefined search terms (e.g., customer preferences/opinions relating to a given product may exist outside of the specific review page of a product). Furthermore, reviews on product review sites are typically generated by customers who purchase the products from such websites. Hence, their reviews can be tainted by experience with the service that such websites provide, not purely on the quality of the products themselves. For

Contributed by the Computers and Information Division of ASME for publication in the Journal of Computing and Information Science in Engineering. Manuscript received June 20, 2014; final manuscript received December 11, 2014; published online xx xx, xxxx. Assoc. Editor: Joshua D. Summers.

example, there are several reviews that leave negative feedback to the products due to slow shipment, dead-on-arrival, poor customer service, etc., instead of reviews about the products themselves.

(3) Bias: Recent research has identified that product review sites such as Amazon.com can be used as channels for companies or interest-sharing third parties to spread spam reviews that persuade customers toward purchasing their products, or dissuade them to shy away from their competitors' products [17,18].

(4) Accessibility: Most social media companies provide tools to easily access full or partial information generated by their users. On the other hand, data existing in web-based content (e.g., customer review data) may be more difficult to extract, requiring a manually created adhoc web-crawler for a different website [19].

(5) Heterogeneity: Compared to web-based content (e.g., product reviews), social media provides its users with the flexibility of expression, resulting in a wide variety of opinions [20]. This heterogeneity in content of social media hence provides an opportunity for users to express opinions about products outside the review sites, especially opinions and expectations toward products or product features not yet existing in the market space.

Social media services such as Twitter and Facebook can be referred to as "digitized word of mouth" as they enable effective, seamless communication by allowing one's opinion to be perceived by a diverse audience [21]. Being ubiquitous and colloquial in nature makes social media a large-scale, upto-date source for mining useful opinions from its users. Most social media providers offer application programming interfaces (APIs) for asking permissions for user data access, hereby providing a seamless means for acquiring large amounts of data in an automated manner. In addition, social media users typically express their personal opinions/preferences publicly, even during product use. For example, messages such as: "I LOVE MY NEW GALAXY S 4G" and "Rip Galaxy 4G:(:(:(:(:(" are common. Knowing that one individual likes/dislikes a particular product or product feature may not be interesting, but millions of such messages may reveal desired product/product features. While many studies have analyzed social media in a wide range of emerging applications, research into the use of social media data to mine product attributes relevant to customers' purchasing decisions (prior to launch and during product usage) has been limited.

A product may come with a *strong* feature that satisfies a majority of customers' needs as well as a *weak* feature that is undesirable to most customers. The ability to automatically identify successful and failing products along with their strong and weak product features could enable designers to refine next generation product designs prior to launch and hence, increase the probability of market success. We propose a methodology that mines product related information from social media data to help designers finetune the features of next generation products. The methodology is based on sentiment analysis and natural language processing techniques that models customers' perception on products in order to understand the factors (e.g., particular product features positively/negatively perceived by customers) that may potentially lead to dissatisfied customers and product returns. Specifically, the methodology has two main components:

(1) Identifying successful/failed products using customers' perception expressed through social media data. We propose to model customers' favorable attitude toward a product by mining their sentiments expressed through social media. Such a measure could be used to predict a product's *Favorability*, or the ability to maintain its impression on customers over time. Previous methodologies quantify overall customers' (positive) perception toward a product by simply aggregating the review scores. However, such an approach can be biased in two senses: (1) Good products which receive only a small number of reviews can be under-valued. (2) Poor or average products with high positive reviews when they are newly launched due to fake reviews and buzz can be over-valued. We address these two issues and frame this prediction problem into a ranking problem where each product is given a *product favorability* (PF) score used to determine its long term ability to remain favorable in the market space.

(2) Identifying product features and opinions consistent with successful/failed products. We introduce a technique to retrieve relevant product features, comprised of *strong* and *weak* features. We further extract *customer opinions* associated with each feature. Such insights could help designers understand why certain products in the market are successful, while others are an abysmal market failure, and help designers develop innovative features for next generation products to satisfy future market needs.

The rest of this paper is organized as follows. Section 2 outlines the literature most closely related to this work. Section 3 discusses the proposed methodology used to address the two challenges outlined above. Section 4 introduces the case studies along with the experimental results and discussion. Section 5 concludes the paper.

2 Background and Related Work

Literature on knowledge-based systems that aid product design and development is extensive [22–24]; however, work pertaining to potential usages of knowledge from social media data in such applications is limited. Hence, this section only discusses previous works closely related to this research.

2.1 Identifying Relevant Product Features and Associated Opinions. Quantifying customer preferences toward different product features may enable designers to understand the aspects of a product that lead to negative customer experiences and ultimately, returned products. Lim et al. proposed a Bayesian network for modeling user preferences on product features [25]. The model is capable of expressing the uncertainty toward product features, and takes into account a user's distribution of preferences over all features. A case study of four laptop product lines shows that their approach was successful in analyzing in-depth component and platform impact under drifting preferences. Tucker and Kim proposed a machine learning based approach for mining product feature trends in the market from the time series of user preferences [15]. 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 [26] and identifying relevant product features from a high dimensional feature set [27]. Ghani et al. proposed a method for identifying product feature-value pairs from textual data [28]. Similarly, Putthividhya and Hu proposed a bootstrapping algorithm for identifying product features and values from online listings [29]. Their methods, however, rely on predefined dictionary of features and attribute values, while our proposed algorithm can extract features unknown to the system. Popescu et al. presented *OPINE*, an unsupervised system for extracting product features from user reviews [30]. For a given product and a corresponding set of reviews, the system is able to extract features along with opinions of the users toward particular features. They used seven product models along with their corresponding web-based reviews for the experiment. Such methods rely on the completeness of the content and correct use of language, and would fail to capture product features discussed in social media where colloquialness and noise are prevalent. Furthermore, most of the above techniques utilize the data from product review sites, whose content pertains to products recently purchased, as opposed to content pertaining to product usage over time. The proposed methodology in this paper aims to model customer product preferences during actual product usage in order to quantify the temporal changes in customer preferences and identify unfavorable/favorable product features that can help guide next generation product designs. Therefore, existing techniques particularly designed for handling data from product review sites are not well suited.

2.2 Social Media as a Viable Modeling Platform. Building knowledge-based systems using useful information from social media data has been extensively studied [14,31]. Acting as a digital word of mouth network makes social media a viable means of spreading content knowledge, which may affect the decision making process of the end users. With this knowledge, one could predict the outcomes of certain events by observing the behaviors emitted from social media. Asur et al. successfully used tweets collected during a three month period to predict box office revenues [32]. They showed that the prediction results were more accurate than those of the Hollywood Stock Exchange. Bollen et al. defined seven dimensions of public moods namely *Calm*, *Alert*, *Sure*, *Vital*, *Kind*, and *Happy* [13]. They modeled the changes of such moods on tweets collected during a 10 month period in 2008, and showed that the changes of such moods correlate with the shifts in the Dow Jones Industrial Average that occur 3–4 days later.

While social media data have been used to model and predict real world phenomenon, product design research pertaining to product feature mining has primarily focused on customer review data, as opposed to social media data [33]. Given the veracity of social media data in predicting real world events, we aim to develop predictive models that help designers understand the factors that influence customers' dissatisfaction/satisfaction when using products.

3 Methodology

We leverage the potential design knowledge existing within social media data to quantify the ability of products to satisfy customers' needs. The mathematical models introduced in this work will also enable designers to determine the set of product features to be incorporated or excluded from next generation products. First, the social media data is collected and preprocessed by removing possible nonhuman generated messages and quantifying levels of sentiment. Note that colloquial content is not removed from the social media data in the preprocessing step, since the authors have shown in previous work that cleaning social media is nontrivial and comes at the risk of losing potential relevant information [31]. The methodology then mines relevant information from the preprocessed data to help designers make crucial decisions regarding design, development, and manufacturing of their future products.

Figure 1 illustrates our proposed methodology that begins with a set of existing products to be explored for relevant product features. These products may include previous product models in the same line or competitors' products. Next, the *Favorability* score, representing customers' long term favorable attitude toward a product, is calculated for each product. The products are then ranked by the *Favorability* scores, and only the top (most favorable) and bottom (least favorable) K products are chosen as *base products*. A base product is an existing product whose notable features can be potentially integrated into next generation products. Only top and bottom K base products are chosen because special consideration should be made for products that satisfy (fail to satisfy) customers' needs. For each chosen base product, its notable features and associated user opinions are extracted. Extracting notable product features allows designers to identify strong and weak features of the existing products. If the base product satisfies customers' needs during product use (characterized by a high PF score), then special consideration is made toward incorporating its

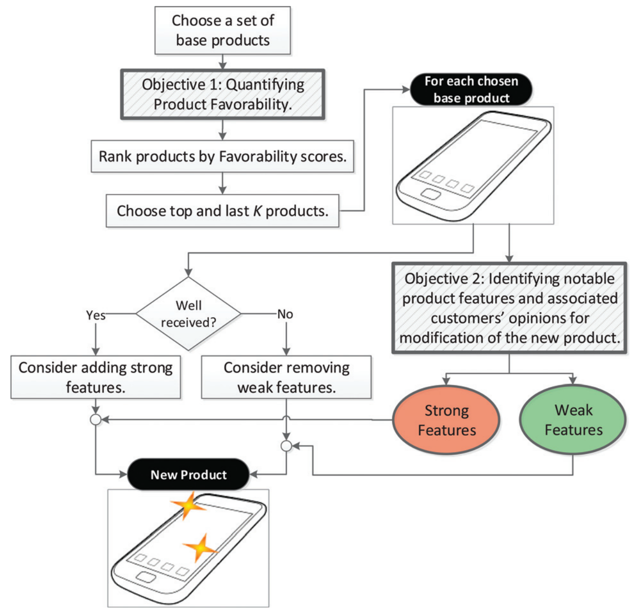

Fig. 1 Proposed method for quantifying PF and product features

strong features in next generation product design efforts, since it is more likely that such a product tends to have favorable and reliable features than those that cause customer dis-satisfaction. On the other hand, if the base product is poorly received (characterized by a low favorability score), then designers should consider removing these weak features when designing the new product, as their inclusion may lead to higher customer dissatisfaction and product returns.

Once the appropriate features are synthesized into the next generation product, designers can then announce the prototype of the next generation product on social media outlets, which would further be discussed among interested customers. Designers may measure the volume of demand toward the new prototype by utilizing social media to predict demand, ahead of product launch [34]. If the new prototype is in high demand, then the company may continue to keep the product in the market space; otherwise, it may choose a new set of base products and repeat the process. The two main components (as shown in bold-gray *objective* boxes in Fig. 1) are proposed and comprehensively investigated in this work. The first objective investigates the possibility of using social media to quantify customers' favorability toward an existing product. The second objective mines social media data in order to discover notable product features.

For designers, the sooner they know what features drive a product to success or failure, the sooner they can design future products that better suite rapidly evolving market needs, potentially providing a competitive advantage in a highly competitive market.

3.1 Social Media Data Collection and Preprocessing. For generalization, the proposed methodology minimizes the assumption about functionalities of social media data, and only assumes that a unit of social media data is a tuple of unstructured textual content 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, Google+, etc. Social media messages corresponding to each product domain are retrieved by detecting the presence of the product's name (and variants).

Social media messages conveying information about products can be divided into two categories: *Product Specification*

ID: veeraragavanb Time: 20:58 I Path: W:/3b2/JCIS/Vol00000/150006/APPFile/AS-JCIS150006

messages and *Product Opinion* messages. *Product Specification* messages objectively describe the features of a particular product, while *Product Opinion* messages express opinions (positive, negative) relating to a particular product/product feature. Listed below are some examples of product specification and product opinion related social media messages about the *Apple iPhone 4* features:

Product Specification: `Closest thing to a retina display computer monitor… the IBM T221 (from 2005) was 22′ WQUXGA (3840?2400). That's 204ppi, iPhone4 = 326ppi.`

Product Opinion (Positive): `Absolutely loving my new iPhone 4 (p.s. I wrote this tweet with #siri lol)`

Product Opinion (Negative): `I hate the fact that my iPhone 4 home button is intermittently unresponsive.`

Social media holds sentiments expressed by its users toward a product. By examining a large number of social media messages relating to product features, it is observed that *Product Opinion* messages usually insinuate emotion of the customers. With such knowledge, we utilize user sentiments in social media to discover individual preferences toward particular products and product features. 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 [35]. The first value represents the *positive* sentiment level, and the other represents the *negative* sentiment level. If a product related message has dominant positive/negative sentiment, it is assumed that the poster likes/dislikes particular features of the product. 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 [36].

In this work, we are primarily concerned about the overall sentiment of a message; hence the positive and negative scores are combined to produce a single emotion strength score using the following equation:

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

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 [37]. 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), 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.

3.2 Objective 1: Quantifying PF Scores. Successful products tend to have good features that impress customers over time, as reflected in both high activity discussion and lasting impressions expressed by customers, measured at the present time. Such ability is defined in this work as the PF. This section introduces a mathematical model that incorporates sentiment in social media messages pertaining to a particular product to calculate the PF score.

Customer satisfaction toward a product has been approximated using the average customer review score already available on product review sites [38]. However, such a method which utilizes product ratings available on review websites can be biased in the following ways:

(1) Fad Products: A recent study has shown that some products can be short lived, but have large amounts of positive reviews [17]. The positive reviews of these products are usually intentionally generated by the companies or interest-sharing parties to boost product sales and attention from customers. Hence, these products tend to be popular for short time before fading away from the market space. Aggregating review ratings of these *fad* products may take those *spam* ratings into account and hence over-value the customer long term satisfaction.
(2) Nonpopular Products: Some products with good features may be known by a few people, resulting in good but few reviews. These products can be under-valued by the traditional satisfaction quantification method.

We reduce such biases by using the information from social media, where users constantly produce messages complaining/admiring products or product features during product usage. Furthermore, unlike traditional consumer satisfaction quantification methods that only take *Popularity* into account, our PF scoring function also considers the *Polarity* and *Subjectivity*, which altogether can characterize the long term customer impression of the product. The subsequent sections explain these measures in detail.

Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of n products and $Positive(s_i)$/$Negative(s_i)$/$Neutral(s_i)$ (refer to Sec. 3.1) be the set of +ve/−ve/0ve messages corresponding to the product s_i.

3.2.1 Polarity. *Polarity* quantifies the long-term impression on a particular product. Products with favorable product reviews tend to satisfy the customers' needs for a long period of time, as reflected by long term customers' polarity (negative or positive opinions) toward the products. For example, the ability to automatically sync content such as music and movies from ITUNES[1] software makes the *iPhone*s appealing for users who regularly listen to music or watch movies from ITUNES. Such impressiveness of a product's features can be captured using the sentiment in social media messages, defined here as *Polarity*:

$$\text{Polarity}(s_i) = \frac{|\text{Positive}(s_i)|}{|\text{Positive}(s_i)| + |\text{Negative}(s_i)|} \quad (2)$$

The notion of *Polarity* in the social media domain is first used in Ref. [32] and is modified here so that the range is between 0 and 1, for consistency when combining with the other components.

3.2.2 Subjectivity. However, good features alone do not make customers satisfied for an extensive period of time. Competitors work hard to make comparable or better features. For example, Blackberry Messenger (BBM)[2] allows Blackberry phone users to send messages to each other over WiFi without the need of texting plans. Shortly thereafter, however, WhatsApp Messenger[3] was developed as an iPhone app to not only include the BBM features, but also add more/better functionality such as the ability to send messages, photos, voices across different mobile platforms. As a successful result, WhatsApp has over 250 million monthly active users (as of June, 2013), while BBM has only 60 million monthly active users (as of May, 2013), despite being on other platforms other than Blackberry (e.g., BBM for Google's Android mobile platform).[4]

Hence, it is also important that the features enabling a product to satisfy customer needs in the market must also be *new and distinct*, that make such a product relevant. Fortunately, new and distinct features usually occur with a lot of diverse discussions about the pros and cons. The volume of controversial discussion about product features is captured by the *Subjectivity*, defined as

[1] http://www.apple.com/itunes
[2] http://us.blackberry.com/bbm.html
[3] http://www.whatsapp.com/
[4] http://www.firstpost.com/blogs/what-bbm-on-android-ios-will-have-that-whatsapp-doesnt-1098791.html

ID: veeraragavanb Time: 20:58 I Path: W:/3b2/JCIS/Vol00000/150006/APPFile/AS-JCIS150006

$$\text{Subjectivity}(s_i) = \frac{|\text{Positive}(s_i)| + |\text{Negative}(s_i)|}{|\text{Positive}(s_i)| + |\text{Negative}(s_i)| + |\text{Neutral}(s_i)|} \quad (3)$$

The notion of *Subjectivity* in the social media domain is first used in [32] and is modified here so that the range is between 0 and 1, for consistency when combining with the other components.

3.2.3 Popularity. Good and newly distinct features may keep customers satisfied. However, a product may not succeed in the market if it is popular among only a few people. For example, *Kyocera Echo*'s notable features include a sturdy body, dual touch screens, and predictive text input. In fact, the user reviews, if any, of the product are mostly positive (4/5 stars by 13 user reviews on Amazon.com,[5] 3.5/5 stars (Very Good) on CNET Editors' Rating,[6] etc.). However, it is hard to find such a smartphone model in the market at the present time, leading many to believe that it has been discontinued by the designer. Not surprisingly, the *Kyocera Echo* page on a popular smartphone review site[7] has a total of only 48,372 views (compared to a successful model such as *iPhone 4*, which has total views of 16,199,129). Hence, the capability of being known and liked by a large group of people should be taken into account when computing the *Favorability*. The *Popularity* score quantifies this

$$\text{Popularity}(s_i) = \frac{|\text{Positive}(s_i)| + |\text{Neutral}(s_i)|}{\sum_{s \in S} (|\text{Positive}(s)| + |\text{Negative}(s)| + |\text{Neutral}(s)|)} \quad (4)$$

The *Popularity* score is normalized to [0,1] range for consistency when combining with the other components.

3.2.4 PF Score. The PF score is computed by combining the three aspects described above which contribute to the long-term product satisfaction, and is defined as

$$\text{PF}(s_i) = \text{Polarity}(s_i) \times \text{Subjectivity}(s_i) \times \text{Popularity}(s_i) \quad (5)$$

$PF(s_i)$ returns a real number between 0 and 1, and is served as a comparative score for ranking products in the same domain, instead of an absolute score. Note that the additive model with each component carrying equal weight was explored but the multiplicative model allows the scores to be more discriminative and suitable for ranking. Such multiplicative models (e.g., *Term Frequency-Inverse Document Frequency (TF-IDF)* and its variants) are widely used in the information retrieval field to rank search results [39]. Additive models with each component carrying a different weight could be explored; however, since the scores are aimed to serve as comparative scores (as opposed to absolute scores where weighted additive models would be more appropriate) and parameter weight tuning is not a focus of this research, the multiplicative model is used to combine the three measures.

3.3 Objective 2: Identifying Notable Product Features.

This section proposes an approach to mine notable features of a product from social media messages that discuss it, and is corresponding to *Objective 2* in Fig. 1. Messages about a product can infer some information about the product features. For example, "FaceTime Iss Amazing:) #iPhone4" implies that the poster likes the *FaceTime* feature of the *iPhone 4*. Similarly, "I hate the iphone 4 battery it keeps dyingUghh" infers that the poster is not satisfied with the *battery life* of her *iPhone 4*. The ability to automatically identify the strong and weak features of a product from the user perspectives could prove to be useful for designers and enterprise decision makers when designing next generation products. Multiple algorithms have been proposed in the literature to extract product features from textual data [30,40]; however, these algorithms would not be applicable in our research due to the reliance on the following assumptions:

(1) Each piece of textual data (i.e., a message) is grammatically correct and rich in textual content. These properties do not hold true for social media data where sparsity and noise are norms.
(2) Each message contributes to discussing product features. However, social media discussion is diverse in topics, some of which relate to product features. A message that mentions a product name does not always discuss about its features.

Not surprisingly, these published algorithms were tested on product review data on which the above assumptions hold. In fact, we tried the algorithm proposed in Ref. [40] and results were full of noisy terms unrelated to product features. In this work, we proposed a new approach to extract *strong* and *weak* product features from sparse and noisy textual data. Strong features make the product appealing to the customers, while weak features make it undesirable. A feature is defined as a noun term representing a property of a product. For example, features for smartphones include *screen*, *app*, *camera*, *battery-life*, etc.

Messages related to a product are divided into +ve, −ve, and 0ve groups. Each message is preprocessed by lowercasing and removing the product names, hashtags, usernames, and punctuation. All terms in the message content is tagged with part of speech (POS) using the Carnegie Mellon ARK Twitter POS Tagger[8] [41], and only noun terms are chosen. A preprocessed message is then composed of a mixture of noun terms representing potential product features.

The feature extraction problem is transformed into the *term ranking* problem, which is then solved using existing information retrieval techniques. For consistency with the information retrieval literature, a message is said to be a *document*. A document d is a bag of terms $T = \{t_1, t_2, \ldots, t_n\}$. Given a set of documents $D = \{d_1, d_2, \ldots, d_m\}$, subset $\theta \subseteq D$, the term ranking algorithm takes the following steps:

Step 1: The set of all distinct terms T are extracted from D.
Step 2: For each term $t \in T$, compute $P(t|\theta, D, T)$, the likelihood (relevant to product features) of the term t given θ, D, and T.
Step 3: Rank the terms by their likelihood.

The above algorithm processes a set of messages corresponding to a product and produces relevant features (represented by noun terms) of the product. As mentioned above, social media users engage in diverse discussion, which may not be related to product features. To mitigate this issue, we first model topics from the set of social media messages, then select topics relevant to product features to compute $P(t|\theta, D, T)$.

Let *Positive*(s)/*Negative*(s)/*Neutral*(s) be the sets of +ve/−ve/0ve tweets related to the product s. The positive/negative features of the product s are the top ranked terms returned by the term ranking algorithm where $D = \text{Positive}(s) \cup \text{Negative}(s) \cup \text{Neutral}(s)$ and $\theta = \textit{Positive}(s)/\textit{Negative}(s)$, respectively.

The next subsections introduce the *latent Dirichlet allocation* (LDA) algorithm which we use to model topics and discuss product feature extraction in detail. AQ1

3.3.1 Topic Modeling With LDA. In text mining, the LDA [42] is a generative model that allows a document to be represented by a mixture of topics. Past literature such as Ref. [31]

[5]http://www.amazon.com/Kyocera-Echo-Android-Phone-Sprint
[6]http://reviews.cnet.com/smartphones/kyocera-echo-sprint/4505-64527-34498252.html
[7]http://www.gsmarena.com/kyocera_echo-3914.php
[8]http://www.ark.cs.cmu.edu/TweetNLP/

demonstrates successful usage of LDA to model topics from given corpora.

The intuition of LDA for topic modeling is that an author has a set of topics in mind when writing a document. A topic is defined as a distribution of terms. The author then chooses a set of terms from the topics to compose the document. With such assumption, the whole document can be represented using a mixture of different topics. LDA serves as a means to trace back the topics in the author's mind before the document is written.

Borrowing the intuition from the original LDA applications, we instead treat a document term as a potential product feature. Therefore, a social media message is instead composed by a mixture of product features. Modeling a document with topic distribution provides the capability to identify whether a document is discussing about product features, by measuring the relevance of the product feature related topics assigned to the document. For example, "I could really use a 5th row of apps on my iPhone 4S home screen.:)" would have high distribution on product feature related topics since the message conveys information about the *app* and *home screen* features of the *iPhone 4S*.

Mathematically, the LDA model is described as following:

$$P(t_i|d) = \sum_{j=1}^{|Z|} P(t_i|z_i = j) \cdot P(z_i = j|d) \tag{6}$$

where $t_i \in T$ and $d \in \theta$. $P(t_i|d)$ is the probability of term t_i being in document d. z_i is a latent (hidden) topic. $|Z|$ is the number of topics. $P(t_i|z_i = j)$ is the probability of term t_i being in topic j. $P(z_i = j|d)$ is the probability of picking a term from topic j in the document d.

The LDA model is used to find $P(z|d)$, the topic distribution of document d, where each topic is described by the distribution of term $P(\mathbf{T}|z)$. Five topics are modeled from θ. In order to identify product feature related topics, two topics whose highest numbers of feature terms within the first 30 terms ranked by $P(t|z)$ are chosen. Two topics are chosen because not all the topics are relevant to product features. The term distribution of the two chosen topics is averaged to represent the new term distribution of the merged topic z^*. Finally $P(t|\theta, D, T)$ can be directly taken from the distribution of the merged topic z^*:

$$P(t|\theta, D, T) = P(t|z^*) \tag{7}$$

3.3.2 Mining Customer Opinions Related to Product Features. Knowing that a product feature is preferable or undesirable could help designers to drill down into specific parts of the product to make adjustments. However, it does not provide much detail on *how* the adjustments should be made. For example, knowing that customers have negative sentiment toward the *video* feature of a smartphone product is not very informative when it comes to actually synthesizing the feature (i.e., what has to be done to improve the video feature). However, knowing that the *video* feature is undesirable because it is perceived as being *slow* and *low-resolution* could potentially help designers to pin-point into what needs to be done to make necessary improvements.

In this section, a natural language processing based approach that utilizes the bootstrapping learning algorithm [43] to extract feature-opinion mappings about product features from a large collection of social media messages is explored to provide more information about what customers *think* toward product features (rather than just negative and positive). The algorithm is also able to extract sample messages which prevalently emit such opinions. These sample messages could be useful for designers to drill down into what actually is said about a feature-opinion pair.

Algorithm 1: The feature-opinion mining algorithm from a collection of social media messages

Input: D(s): Set of social media messages related to product *s*.
F: Set of features

Output: E: Set of extractions. Each $e \in E$ is a tuple of $\langle feature, opinion, \{relevantmessages\}\rangle$, for example
$e = \langle \text{'}onscreen\ keyboard\text{'}, \text{'}fantastic\text{'}, \{d_1, d_2, \ldots\}\rangle$

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

The opinion mining algorithm used in this paper was first developed by Huang et al. to mine opinions related to restaurants in Seattle area from Yelp reviews [40]. The algorithm was later modified by Tuarob and Tucker so that it could handle noisy data such as social media data more efficiently [6]. The modified algorithm is outlined in Algorithm 1. The input is a collection of social media messages related to product s, $D(s)$. The algorithm then preprocesses each message by cleaning residuals such as symbols, hyperlinks, usernames, and tags, correcting misspelled words, and removing artificial generated messages. Such noise is ubiquitous in social media and could cause erroneous results. The Stanford POS Tagger[9] is used to tag each word with an appropriate POS. This step is required because the template learning algorithm relies on the grammatical structure of each sentence, defined by a sequence of parts of speech.

The main part of the algorithm iteratively learns to identify feature-opinion pairs and generates a set of extractions ($E(s)$) related to the product s from the input collection of social media messages. The algorithm employs a bootstrapping learning algorithm where a set of ground-truth features is fed as seed features. The algorithm then repeatedly learns phrase templates surrounding the seed features, and uses the templates to extract more opinions associated with each feature. This process continues until the extraction set does not grow.

Finally, the algorithm postprocesses the extractions by disambiguating and normalizing the opinions. The disambiguation process involves stemming the opinions using the Porter's stemming algorithm[10] and clustering them using the WordNet[11] SynSet. This postprocessing step groups the same opinions, which may be written differently together (e.g., `amazing`, `amaze`, `amazes`, and `fascinating` would be grouped together). The final output is a set of *extractions* where each extraction $e \in E(s)$ is a tuple of $\langle feature, opinion, snippets\rangle$ such as:

⟨**feature**: "`onscreen keyboard`,"
opinion: "`fantastic`,"

[9]http://nlp.stanford.edu/downloads/tagger.shtml
[10]http://tartarus.org/martin/PorterStemmer/
[11]http://wordnet.princeton.edu/

snippets: { "This onscreen keyboard is fantastic with text prediction," "Fantastic! now i can use swipe features on the onscreen keyboard"} ⟩

To illustrate the use of the above example, after the notable feature extraction phrase, designers may find that the *onscreen keyboard* is a strong feature of a competitor's product. Designers would then want to know *why* it is a strong feature. The example opinion mining result above would help explain that some customers view such a feature as *fantastic* due to the compatibility with the *text prediction* and *swipe* features. Designers could use this knowledge to decide whether it is possible to add such capability to their target next generation products.

4 Case Studies

Two case studies (smartphone and automotive products) are presented that use social media data (Twitter data) to mine relevant product design information. Data pertaining to product specifications from smartphone and automotive domains are then used to validate the generated models in the objective components of the proposed system.

4.1 Data Acquisition

4.1.1 Model Generation Data: Twitter Data. Twitter[12] is a microblog service that allows its users to send and read text messages of up to 140 characters, known as *tweets*. The tweets used in this research were collected randomly using the provided Twitter API, and comprises roughly 800×10^6 tweets in the United States during the period of 19 months, from March 2011 to September 2012.

4.1.2 Model Validation Data 1: Smartphone Specification Data. The smartphone database is obtained from GSMArena.[13] GSMArena catalogs a majority of cellphone models along with their technical specification, user rating, and user comments. All the smartphone pages in GSMArena are crawled and parsed to obtain necessary information. The database consists of 2547 smartphone models designed by 33 different companies.

4.1.3 Model Validation Data 2: Automobile Specification Data. Twenty-nine automobile products reported to be the worst and the best by the Consumer Reports[14] magazine published in April 2011[15] are chosen for the case studies. The car ratings are taken from both the Consumer Reports magazine (April 2013)[16] and USNews.com.[17]

4.2 Objective 1: Quantifying PF Scores.

To evaluate the proposed *Favorability* scoring model, 21 smartphone models and eight automobile models are chosen for this analysis. The smartphone models include *Apple iPhone 4*, *Samsung Galaxy Nexus*, *Samsung Galaxy Tab*, *Samsung Galaxy S II*, *Motorola Droid RAZR*, *HTC ThunderBolt*, *Sony Ericsson Xperia Play*, *Motorola DROID X2*, *Samsung Infuse 4G*, *BlackBerry Bold 9900*, *Nokia N9*, *Samsung Galaxy S 4G*, *HP Veer*, *Dell Venue Pro*, *T-Mobile G2x*, *Kyocera Echo*, *Nokia E7*, *Samsung Dart*, *LG Cosmos Touch*, *Samsung Exhibit 4G*, and *LG Enlighten*. The automobile models include *Toyota Camry*, *Toyota Prius*, *Toyota Corolla*, *Honda Civic*, *Nissan Sentra*, *Honda Accord*, *Jeep Wrangler*, and *Nissan Altima*.

[12]https://twitter.com/
[13]gsmarena.com
[14]Consumer Reports is an American magazine published monthly by Customers Union since 1936. It publishes reviews and comparisons of customer products and services based on reporting and results from its in-house testing laboratory and survey research center. It also publishes cleaning and general buying guides.
[15]http://www.customerreports.org/cro/magazine-archive/2011/april/april-2011-toc.htm
[16]http://www.customerreports.org/cro/magazine/2013/04/
[17]http://usnews.rankingsandreviews.com/cars-trucks

Table 1 Numbers of positive, negative, neutral, and all tweets related to each selected smartphone model

Model\# Tweets	# Pos	# Neg	# Neu	# All
iPhone 4	29013	15657	50362	95032
Samsung Galaxy Nexus	1330	698	2284	4312
Samsung Galaxy Tab	946	432	1762	3140
Samsung Galaxy S II	1021	438	1643	3102
Motorola Droid RAZR	578	300	886	1764
HTC ThunderBolt	332	173	537	1042
Sony Ericsson Xperia Play	102	51	249	402
Motorola DROID X2	99	58	214	371
Samsung Infuse 4G	91	34	143	268
BlackBerry Bold 9900	96	27	133	256
Nokia N9	64	30	91	185
Samsung Galaxy S 4G	54	25	93	172
HP Veer	44	20	77	141
Dell Venue Pro	39	8	35	82
T-Mobile G2x	27	6	47	80
Kyocera Echo	13	10	27	50
Nokia E7	7	5	13	25
Samsung Dart	6	6	10	22
LG Cosmos Touch	8	1	9	18
Samsung Exhibit 4G	6	1	10	17
LG Enlighten	3	0	14	17

Table 2 Numbers of positive, negative, neutral, and all tweets related to each selected automobile model

Model\# Tweets	# Pos	# Neg	# Neu	# All
Toyota Camry	5440	2168	6023	13631
Toyota Prius	4328	3582	6858	14768
Toyota Corolla	1756	1017	3796	6569
Honda Civic	1704	942	2505	5151
Nissan Sentra	949	534	1562	3045
Honda Accord	839	427	1344	2610
Jeep Wrangler	643	329	1043	2015
Nissan Altima	406	157	746	1309

Tables 1 and 2 break down the numbers of positive, negative, neutral, and all tweets corresponding to each smartphone and automobile model, respectively.

For smartphone products, the *Favorability* scores are computed for the 21 smartphones. The scores are compared with the GSMArea's *Daily Interest* ratings. The ratings from GSMArea are used as ground truth validation data due to the reliability of the websites along with the availability of the data for all the chosen 21 smartphone models. The Daily Interest rates used here are the average of three consecutive days starting from January 4, 2013. Figure 2 plots the normalized *Favorability* scores against the normalized GSMArena ratings in log scale. The log scale is used to illustrate the ability to produce rankings for products with low reputations (whose scores converge to near zero). A high ranking correlation of 0.8182 is observed between the rankings produced by *Favorability* scores and the GSMArea Daily Interest rates. Since all the 21 smartphone models were released in 2011 or before, the ability to satisfy current customer needs with such models is reflected in current interest levels expressed by current customers, supporting the high correlation found.

For automobile products, the *Favorability* scores are computed for the eight automobile models. The user ratings from the U.S. News Car Ranking and Reviews 2013[18] and Consumer Reports (April 2013) ratings are used as ground truth validation data. The ratings are used to reflect today's interest on the selected automobile products. Figure 3 plots the normalized results. High ranking correlations of 0.7857 and 0.9524 are observed between the

[18]http://usnews.rankingsandreviews.com/cars-trucks/

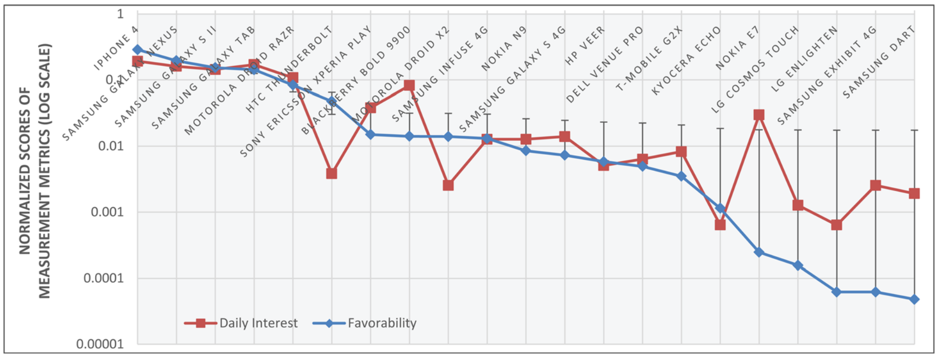

Fig. 2 Comparison between the PF score versus GSMArea daily interest for each sample smartphone model (in log scale). The products are ordered by their *Favorability* scores.

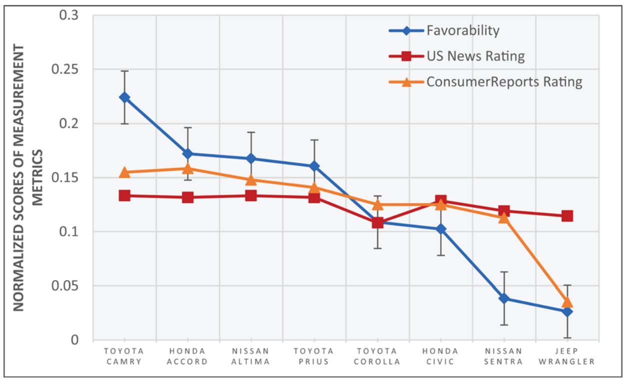

Fig. 3 Comparison between the Favorability score versus U.S. News and consumer reports ratings for each sample automobile model. The models are ordered by their *Favorability* scores.

rankings produced by the *Favorability* scores and the ratings from the U.S. News and the Consumer Reports magazine, respectively.

A natural question would be: why not use these well established scores (e.g., GSMArena and Consumer Reports) directly, instead of computing the *Favorability* scores from social media data? While using the product comparison scores from well-established sources may be an obvious option, it faces the following challenges:

(1) Well established product comparison scores from reliable sources are only available for some product categories. Some popular products such as smartphone and automobiles demand reliable comparison metrics to help customers make decision; however, it would be difficult to find reliable comparison scores for some products such as particular dishes in a restaurant or soda beverages in supermarkets. These relatively mundane products are discussed in social media, and hence, it would be possible to compare them directly using the proposed *Favorability* scores.

(2) Well established product comparison sources only allow a small number of products to be compared. For example, U.S. News Car Ranking provides rankings for only 40 automobile products in the "Affordable Small Cars" categories. Hence, the comparison to other automobile products outside this set would be inapplicable.

Designers could use the *Favorability* scores to identify successful and failing products to be used as the *base* products, according to Fig. 1.

4.3 Objective 2: Identifying Notable Product Features. This section reports the results from applying the proposed feature extraction methodology on the Twitter data corresponding to the smartphone and automobile products.

4.3.1 Extracting Notable Product Features. In terms of quantifying notable product features expressed through social media (i.e., Twitter in this case study), we have focused only on products of which specific features expressed in the sample data are available. Four smartphone (*Apple iPhone 4*, *Samsung Galaxy S II*, *Motorola Droid RAZR*, and *Sony Ericsson Xperia Play*) and four automobile products (*Toyota Prius*, *Tesla Model S*, *Honda Civic*, and *Jeep Wrangler*), which have large amount of corresponding tweets, are chosen for the study. In the experiment, the topics are modeled using Stanford Topic Modeling Toolbox[19] with 3000 maximum running iterations and using the collapsed variational Bayes approximation to the LDA objective [44].

Note that the top terms returned by the term ranking algorithm may include random noun terms not relevant to product features. The evaluation in terms of meaningfulness is performed on each ranked list of the 50 terms, using Precision@50 protocol defined as

$$\text{Precision@50} = \frac{|\text{Feature Terms in Top 50 Terms}|}{50} \tag{8}$$

Precision is an evaluation metric extensively used to evaluate a classification system for its ability to retrieve correct objects from a pool of random objects [45]. This score can also be used to interpret the users' knowledgeability about and the richness of the features of a particular product. Products with many notable features tend to urge users to discuss about them, resulting in high volume of discussions related to the product features.

Tables 3 and 4 list the top ten strong and weak features of the chosen smartphone and automobile products respectively, along with the Precision@50 scores. The top ten strong/weak features extracted from the chosen models provide useful information that matches with the actual product specification. Note that if a feature is both strong and weak, then it is a *controversial* feature. A controversial feature is characterized by diverse opinions, leading to both pro and con discussions.

For smartphone examples, the *Apple iPhone 4* features 5 MP and dual (back and front) cameras, longer battery life compared to the predecessor, Retina screen, FaceTime, iMessage messaging system, and Voice Control command. However, some users still complain about the battery time while on 3G mode, harder to jailbreak, and the bug about occasional signal drop when touching the antenna sideline. Note that the features extracted from social media are subjective to social media users; hence, *harder to jailbreak* may be considered a weak feature to the user (who wishes to jailbreak his/her phone), though it might be considered a strong feature from the manufacturer's point of view. Similarly, the *Sony*

[19]http://nlp.stanford.edu/software/tmt

J_ID: JCIS DOI: 10.1115/1.4029562 Date: 19-January-15 Stage: Page: 9 Total Pages: 13

PROOF COPY [JCISE-14-1216]

Table 3 Features extracted from tweets related to each selected smartphone model

	iPhone 4		Samsung Galaxy S II		Motorola Droid RAZR		Sony Ericsson Xperia Play	
Features	Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak
1	Camera	Battery-life	Touch-screen	Touch-screen	Battery-life	Keys	Game	Game
2	Battery-life	Face-time	Update	Function	Screen	Price	Battery-life	Accessories
3	Screen	App	Battery-life	Email	Picture	Browser	Control	Video
4	App	Video	Screen	Video	Android	Bootloader	Fun	Battery-life
5	Price	Jail break	Ics	Bootloader	Glass	Warranty	Hardware	Commercial
6	Music	Wifi	Sensation	Photo	App	Microphone	Performance	Style
7	Face-time	Bug	Display	Gallery	Camera	Delay	Experience	Control
8	Message	Charge	Video	Button	Keyboard	Bloatware	Wifi	App
9	Voice-control	Location	App	Texting	Network	Fixes	Video	Size
10	Case	Touch-screen	Picture	Price	Noise	Email	Controller	Carrier
Pr@50	0.62	0.56	0.52	0.1	0.36	0.26	0.38	0.16

Table 4 Features extracted from tweets related to each selected automobile model

	Toyota Prius		Tesla Model S		Honda Civic		Jeep Wrangler	
Features	Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak
1	Gas	Racing	Electric	Charge	Price	Rims	Fun	Tires
2	Mpg	Drag	Charging time	Gear	Coupe	Coupe	Driving	Drive
3	Driving	Commercial	Supercharger	Miles	Miles	Spoiler	Country	Wheel
4	Hybrid	Environment	Sedan	Electric	Details	Driving	Price	Snow
5	Fuel	Feel	Display	Falsehood	Commercial	Bumper	Wheel	Dirt
6	Service	Style	Fun	Sedan	Auto-trans	Race	Size	Park
7	Smooth	Blind spot	Control	Damage	Hatchback	Mileage	Manual-trans	Safety
8	Quiet engine	Discharge	Technology	Touchscreen	Parking	Engine	Off-road	Noise
9	Gadgets	Charging	Looks	Interior	Sports	Backseat	Exploration	Seats
10	Battery	Tire	Luxury	Price	Style	Cheap	Unique	Looks
Pr@50	0.36	0.32	0.38	0.28	0.28	0.26	0.24	0.24

Ericsson Xperia Play features the combination of smartphone and game console. Hence most of its strong features involve gaming. However, the model comes with a bulky look; hence, style and size come up as weak features. As a practical example for designing a new smartphone product, designers could consider adding successful features of the *Apple iPhone 4* such as the dual cameras and the Facetime, while removing weak features of the *Sony Ericsson Xperia Play* such as the bulky look and style.

Likewise, for automobile products, the *Toyota Prius* is known for its innovative hybrid system that allows the engine to achieve high mpg (miles per gallon). However, the model is also known for a bad design that limits visibility in the blind spots, and slow acceleration that drags the car during racing. The *Jeep Wrangler* is well known for its off-road capability; however, customers have complained for the engine noise and uncomfortable seating. Designers could, for example, design a new car that incorporates strong features from the *Toyota Prius* such as the gas saving feature, while removing the weak features from the *Jeep Wrangler* such as the noise and small seating.

The *Pr@50* scores infer how much proportion of the sample social media data related to a particular product is devoted to discussing the product features. The future work could employ this finding to quantify and compare the richness of features across multiple products. In Table 3, one could clearly see that successful products (i.e., *iPhone 4* and *Samsung Galaxy S II*) overall have higher *Pr@50* scores than the inferior products (i.e., *Motorola Droid RAZR* and *Sony Ericsson Xperia Play*). Though such distinction is not clear in automobile products (according to Table 4), one could observe that the *Jeep Wrangler*, regardless of its unique off-road capabilities, overall has fewer features than the *Toyota Prius* and *Tesla Model S*.

To further validate the extraction of the notable features, the synthesis of features of two smartphone product lines are investigated, including the *iPhone* and the *Samsung Galaxy*. Figures 4 and 5 illustrate the feature sentiment levels (positive and negative) associated with some features of the *iPhone* (i.e., *iPhone 4* and *iPhone 5*) and the *Samsung Galaxy* (i.e., *Samsung Galaxy S II* and *Samsung Galaxy S III*) product lines.

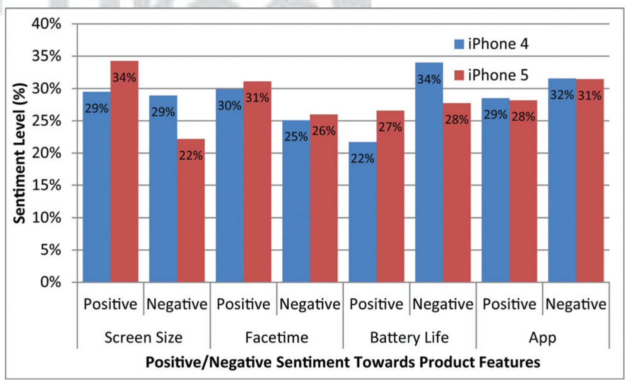

Fig. 4 Comparison between the positive and negative sentiments related to some features of *iPhone 4* and *iPhone 5*

Each positive/negative feature sentiment level of a product feature is calculated by normalizing the aggregate positive/negative sentiment scores of the social messages that mention such a feature of the product. Concretely, for a given feature f of the product s, let $M(s, f)$ be the set of social media messages associated with the product s and mention the feature f. The positive/negative feature sentiment levels $(\mathrm{FSL}^{+}(f,s)/\mathrm{FSL}^{-}(f,s))$ are defined as

$$\mathrm{FSL}^{+}(f,s) = \frac{100\%}{5\cdot|M(f,s)|}\sum_{m\in M(f,s)} \text{Positive Score}(m) \qquad (9)$$

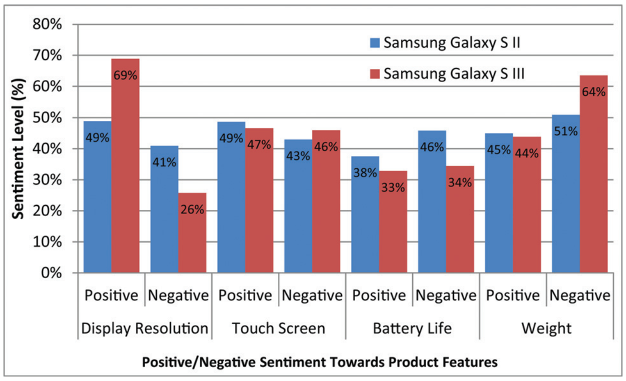

Fig. 5 Comparison between the positive and negative sentiments related to some features of *Samsung Galaxy S II* and *Samsung Galaxy S III*

$$\mathrm{FSL}^{-}(f,s) = \frac{100\%}{5\cdot|M(f,s)|}\sum_{m\in M(f,s)} \text{Negative Score}(m) \qquad (10)$$

Note that the 5 in the denominator is introduced to normalize the positive/negative sentiment scores to the range [0,1] (recall from Sec. 3.1 that a positive/negative score can take the value from 1 to 5).

Each selected product line consists of two products of the consecutive generations (i.e., *iPhone 4* → *iPhone 5* and *Samsung Galaxy S II* → *Samsung Galaxy S III*). Four sample features are selected for each product line including:

FSI: A strong feature of other products outside the product line that is improved in the next generation product.
FSS: A strong feature of the previous product that remains the same or is not improved in the next generation product.
FWI: A weak feature of the previous product that was removed/improved in the next generation product.
FWS: A weak feature of the previous product that remains the same or is not improved in the next generation product.

The strong and weak features are taken from Table 3. For the *iPhone* line, the chosen *FSI*, *FSS*, *FWI*, and *FWS* features are *Screen Size*, *Facetime*, *Battery Life*, and *App*, respectively. Big *screen sizes* have been known as a strong feature of the *Samsung Galaxy* products. Subsequently, the *iPhone 5* has a bigger (longer) screen compared to its predecessor to support another row of apps. Synthesizing this feature turns out to be favorable since the positive FSL increases by 5% while the negative FSL decreases by 7%. The *Facetime* of the *iPhone 5* does not change much (perhaps due to less dependency on hardware). Hence, the positive and negative FSLs remain roughly the same across these two products. The short *battery life* feature was a big complaint in the *iPhone 4*. In the *iPhone 5*, the battery life is extended to 10 hr talk time on 3G (+3 hr, +43%) and 8 hr internet on 3G/LTE (+2 hr, +33%).[20] This battery life extension in the iPhone 5 results in a rise in positive sentiment level by 5% and a drop in negative sentiment level by 6%. The *app* feature is regarded as a weak feature of *iPhone 4*; however, due to being hardware independent, there is no model-specific improvement regarding such a feature. Similar to the *Facetime* feature, the positive and negative FSL remains about equal across the two products.

For the *Samsung Galaxy* product line, the chosen FSI, FSS, FWI, and FWS features are *Display Resolution*, *Tough Screen*, *Battery Life*, and *Weight*, respectively. The high *display resolution* is one of the selling features of the *iPhone 4* which implements a high-resolution Retina display (960 × 640 resolution at 326 ppi). The high-resolution feature is later implemented in the *Samsung Galaxy S III*, which extends the resolution from 480 × 800 pixels at 218 ppi to 720 × 1280 at 306 ppi, bringing the display quality closer to its competitor (still lower ppi compared to the *iPhone*, but more pixels). As a result, the positive FSL rises by 20% and the negative FSL falls by 15%. The *touch screen* feature, though being a weak feature in the *Samsung Galaxy S II*, is not changed nor improved in the *Samsung Galaxy S III*; hence, there is no obvious difference in both the positive and negative FSLs. The *Samsung Galaxy S III* expands the *battery* capacity from 1650 mAh to 2100 mAh, resulting in an extension of the talk time to 22.5 hr (+4.2 hr, 23%) and the stand-by time to 34.6 days (+5 days, 17%). Interestingly, though the negative FSL of the battery life feature decreases by 12% as expected from the improvement, the positive FSL also decreases slightly (only by 5%, however). An explanation for this phenomenon could be that the extension of the battery life in the *Galaxy S III* satisfies the needs from those customers who suffer from the short battery life in the predecessor (judging from the fewer complaints, resulting in lower negative FSL); however, the improvement on the battery life does not extraordinarily impress the customers. This is because, the talk time of the *Galaxy S II* (i.e., 18 hr), which could last more than a day on regular use, is already more than enough for most users who normally charge their phones everyday. Further improving this feature may not be very beneficial for most customers, resulting in nonincreasing positive FSL. The heavy *weight* feature of the *Galaxy S II* is one of the weak features. However, not only is the weight is not reduced in the next generation model, but the *Galaxy S III* is even heavier than its predecessor. This subsequently causes a further rise in the negative FSL by 13%.

These two examples above indicate that incorporating recommended strong features and removing/improving the weak features in the next generation products could increase the overall positive perception among social media users, which may result in higher actual demands for the products in the market space [46,47].

4.3.2 Mining Customer Opinions Related to Product Features. The opinion mining algorithm (Algorithm 1) is applied on the set of social media messages associated with each selected product in the previous section. Recall that the algorithm takes a AQ2
set of social media messages related to a product and a set of product features as input, and outputs opinions and snippets associated to those features. Figure 6 shows an example output from the algorithm on some features (i.e., *case*, *facetime*, and *screen*) of the *iPhone 4*. The algorithm is implemented in JAVA and writes outputs in JSON format which could be further processed in many search and database systems such as JsonEditor[21] and MongoDB.[22] The output is categorized in the hierarchy format of Product Name → Feature → Opinion → Snippets. The snippets are the social media messages that frequently discuss about the product feature ~~(highlighted in blue)~~ and opinion ~~(highlighted in yellow)~~ pair. This model illustrates examples that designers to AQ3
look into what exactly customers discuss about the product features.

Note that not all social media messages that mention a product feature are captured by the opinion mining algorithm. The major reason is because the algorithm cannot find the associated opinions, even though the opinion can be implicitly inferred. Some examples of such messages include "You were *racing*.in a prius? seriously?" (implying the poster might think that Prius is *unsuitable* for *racing*) and "New BlackBerry Bold 9900 with *touch screen*! I want to trade in my Bold for it!" (implying that the new BlackBerry Bold 9900 has *touch screen* that may be superior to the poster's current phone, urging her *desire* to obtain such a phone). Unfortunately, the algorithm is currently unable to detect such implicit semantics; which marks a limitation in this work. Future works could explore techniques such as deep learning for semantic interpretation [48].

[20] http://www.apple.com/iphone

[21] http://www.jsoneditoronline.org/

[22] http://www.mongodb.org/

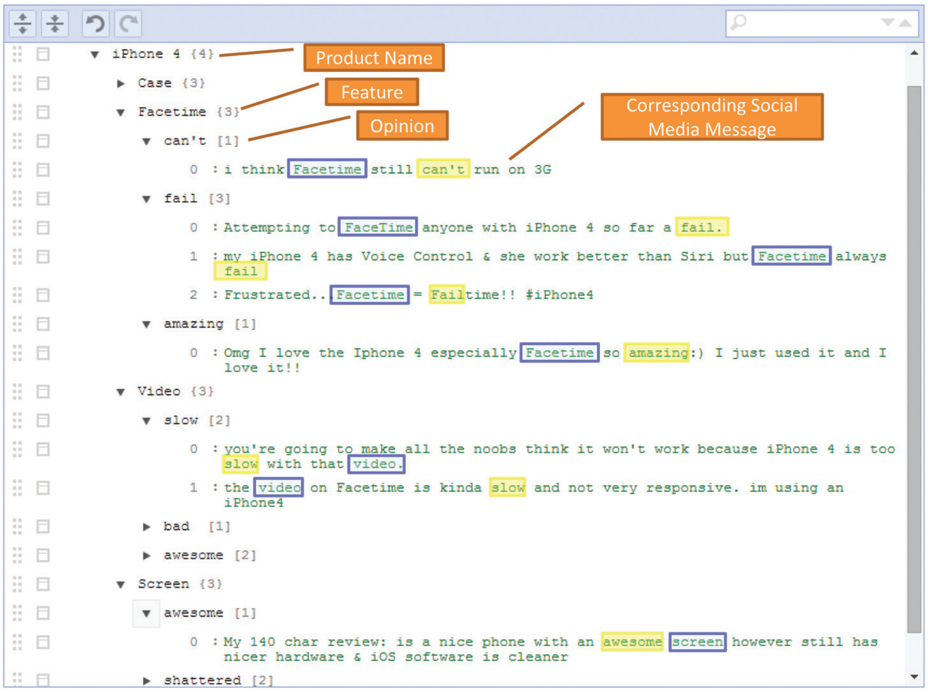

Fig. 6 Sample feature opinions related to the *iPhone 4*, arranged in hierarchy of Product Name → Feature → Opinion → Snippets

Table 5 Top customers opinions, ranked by frequency, related some notable features of *iPhone 4*, *Samsung Galaxy S II*, *Toyota Prius*, and *Tesla Model S*

Model	iPhone 4		Samsung Galaxy S II		Toyota Prius		Testla Model S	
Feature	Camera	Battery-life	Touch-screen	Email	Gas	Racing	Electric	Charge
Opinion 1	Awesome	Dead	Perfect	Slow	Saving	Drag	Working	Few
Opinion 2	Great	Horrible	Big	Horrible	Good	Behind	Awesome	Bad
Opinion 3	Best	Better	Awesome	Blocked	Cheap	Seriously	Complete	Little
Opinion 4	Incredible	Draining	Small	Noticeable	Money	Horrible	Luxury	Rare
Opinion 5	Better	Fixed	Cracked	Connected	Best	Slower	Full	Slow
Opinion 6	Amazing	Empty	Huge	Ugly	Full	Lame	New	Hard
Opinion 7	Bad	Sinking	Vivid	Limit	Expensive	Sick	Great	Reducing
Opinion 8	Like	Decreased	Nice	Loading	Crazy	Limit	100%	Expensive
Opinion 9	Sluggish	Longer	Clear	Fast	Better	New	Expensive	Game-changing
Opinion 10	Cool	Short	Responsive	Okay	Filled	Down	Innovative	Intrigued

Table 5 lists top opinions associated with some features of the *iPhone 4*, *Samsung Galaxy S II*, *Toyota Prius*, and *Tesla Model S*. The extracted opinions are ranked by the frequency of occurrence. Note that the algorithm is run on the entire collection of messages associated with each product; hence, there can be a mix of positive and negative opinions. However, the proportion of positive opinions on *strong* features (i.e., *iPhone 4*'s camera, *Samsung Galaxy S II*'s touch screen, *Toyota Prius*'s gas, and *Tesla Model S*'s electric) are greater than negative opinions. Likewise, the negative opinions of the *weak* features (i.e., *iphone 4*'s battery life, *Samsung Galaxy S II*'s email, *Toyota Prius*'s racing, and *Tesla Model S*'s charge) are more prevalent than the positive ones.

5 Conclusions and Future Work

We proposed a data mining driven methodology that uses large scale data, existing in social media networks to construct a knowledge-based system to support product design and development processes. The system quantifies customers' satisfaction during the usage life of products in an effort to understand the factors that impact customer satisfaction/dissatisfaction. Two main contributions are proposed in this work in an effort to mitigate the wasted design efforts and increased environmental impact that results from returned goods that fail to meet customer needs. The first objective quantifies customer current satisfaction of individual products using their corresponding social media messages, in order to determine strong and weak products. A high ranking correlation between the results from the proposed mathematical model and today's current interest rates from end users is observed. The model is tested on a selection of 21 smartphone and eight automobile products *said* to be the best and the worst in 2011. The second objective employs information retrieval techniques to extract notable (*strong* and *weak*) features and corresponding customers' opinions of individual products from social media. The proposed approach yields promising results that show high correspondence with the actual product features. The extracted notable features could help designers understand why a product performs better or worse than the others, and also help in the design of next generation products.

Designers could use this design knowledge to manage the design and development of their products. Future works could investigate the usage of the buzzes in social media to infer product expectations from customers in order to predict the market reception of product prototypes.

References

[1] Zhang, K., Narayanan, R., and Choudhary, A., 2010, "Voice of the Customers: Mining Online Customer Reviews for Product Feature-Based Ranking," Proceedings of the 3rd conference on Online social networks, USENIX Association, ~~pp. 11–■.~~ AQ4

[2] Richins, M. L., 1983, "Negative Word-Of-Mouth by Dissatisfied Consumers: A Pilot Study," J. Mark., **47**(1), pp. 68–78.

[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] Tuarob, S., and Tucker, C. S., 2014, "Discovering Next Generation Product Innovations by Identifying Lead User Preferences Expressed Through Large Scale Social Media Data," ASME Paper No. 2014-34767.

[7] 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.

[8] Bodnar, T., Tucker, C., Hopkinson, K., and Bilén, S., 2014, "Increasing the Veracity of Event Detection on Social Media Networks Through User Trust Modeling," Proceedings of 2014 IEEE International Conference on Big Data, ■. AQ5

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

[10] 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, (New York, NY), pp. 851–860.

[11] Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H., Mitra, P., Wu, D., Tapia, A., Giles, L., Jansen, B., and Yen, J., 2011, "Classifying Text Messages for the Haiti Earthquake," Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM2011), pp. 1–10. AQ6

[12] Collier, N., and Doan, S., 2012, "Syndromic Classification of Twitter Messages," *Electronic Healthcare*, P. Kostkova, M. Szomszor, and D. Fowler, eds., Vol. 91, Springer, Berlin, Germany, pp. 186–195.

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

[14] Esparza, S. G., O'Mahony, M. P., and Smyth, B., 2012, "Mining the Real-Time Web: A Novel Approach to Product Recommendation," Knowl. Based Syst., **29**, pp. 3–11.

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

[16] Kaplan, A. M., and Haenlein, M., 2010, "Users of the World, Unite! The Challenges and Opportunities of Social Media," Bus. Horiz., **53**(1), pp. 59–68.

[17] Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., and Ghosh, R., 2013, "Exploiting Burstiness in Reviews for Review Spammer Detection," Seventh International AAAI Conference on Weblogs and Social Media, ■.

[18] Chevalier, J. A., and Mayzlin, D., 2006, "The Effect of Word of Mouth on Sales: Online Book Reviews," J. Mark. Res., **43**(3), pp. 345–354.

[19] Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S., 2011, "Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media," Bus. Horiz., **54**(3), pp. 241–251.

[20] Himelboim, I., McCreery, S., and Smith, M., 2013, "Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter," J. Comput. Mediated Commun., **18**(2), pp. 40–60.

[21] Dellarocas, C., 2003, "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," Manage. Sci., **49**(10), pp. 1407–1424.

[22] Fuge, M., Tee, K., Agogino, A., and Maton, N., 2014, "Analysis of Collaborative Design Networks: A Case Study of Openideo," ASME J. Comput. Inf. Sci. Eng., **14**(2), p. 021009.

[23] Yassine, A. A., and Bradley, J. A., 2013, "A Knowledge-Driven, Network-Based Computational Framework for Product Development Systems," ASME J. Comput. Inf. Sci. Eng., **13**(1), p. 011005.

[24] Liu, Y., Liang, Y., Kwong, C. K., and Lee, W. B., 2010, "A New Design Rationale Representation Model for Rationale Mining," ASME J. Comput. Inf. Sci. Eng., **10**(3), p. 031009.

[25] Lim, S. C. J., Liu, Y., and Loh, H. T., 2012, "An Exploratory Study of Ontology-Based Platform Analysis Under User Preference Uncertainty," Proceedings of ASME 2012 International Design Engineering Technical Conference on Computers and Information in Engineering Conference (IDETC/CIE2012), pp. 519–528.

[26] 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.

[27] 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.

[28] Ghani, R., Probst, K., Liu, Y., Krema, M., and Fano, A., 2006, "Text Mining for Product Attribute Extraction," SIGKDD Explor. Newsl., **8**(1), pp. 41–48.

[29] Putthividhya, D. P., and Hu, J., 2011, "Bootstrapped Named Entity Recognition for Product Attribute Extraction," Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP'11, Stroudsburg, PA, pp. 1557–1567.

[30] 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, Stroudsburg, PA, pp. 339–346.

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

[32] Asur, S., and Huberman, B. A., 2010, "Predicting the Future With Social Media," Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT'10, vol. 1, Washington, DC, pp. 492–499.

[33] Wang, L., Youn, B., Azarm, S., and Kannan, P., 2011, "Customer-Driven Product Design Selection Using Web Based User-Generated Content," Proceedings of the 2011 ASME IDETC/CIE, pp. 405–419.

[34] 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 IDETC/CIE2013, ■.

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

[36] Fox, E., 2008, *Emotion Science: Cognitive and Neuroscientific Approaches to Understanding Human Emotions*, Palgrave Macmillan, ■. AQ7

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

[38] Babich, P., 1992, "Customer Satisfaction: How Good is Good Enough?" Qual. Prog., **25**, pp. 65–■. AQ8

[39] Manning, C. D., Raghavan, P., and Schütze, H., 2008, *Introduction to Information Retrieval*, Vol. 1, Cambridge University Press, Cambridge, UK.

[40] 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, New York, NY, pp. 3–12.

[41] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A., 2011, "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments," Proceedings of the 49th Annual Meeting of the ACL: HLT 2011, Stroudsburg, PA, pp. 42–47.

[42] Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003, "Latent Dirichlet Allocation," J. Mach. Learn. Res., **3**, pp. 993–1022.

[43] Thelen, M., and Riloff, E., 2002, "A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pattern Contexts," Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, EMNLP'02, vol. 10, Stroudsburg, PA, pp. 214–221.

[44] Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W., 2009, "On Smoothing and Inference for Topic Models," Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI'09, Arlington, VA, pp. 27–34.

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

[46] Pookulangara, S., and Koesler, K., 2011, "Cultural Influence on Consumers' Usage of Social Networks and Its' Impact on Online Purchase Intentions," J. Retailing Consum. Serv., **18**(4), pp. 348–354.

[47] Ioanăs, E., and Stoica, I., 2014, "Social Media and Its Impact on Consumers Behavior," Int. J. Econ. Pract. Theor., **4**(2), pp. 295–303.

[48] Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y., 2012, "Improving Word Representations Via Global Context and Multiple Word Prototypes," ACL'12, pp. 873–882.