MINING ONLINE PRODUCT REVIEWS TO IDENTIFY CONSUMERS' FINE-GRAINED CONCERNS

Jian Jin[1,2]*, Ping Ji*[2,3]

[1]*Department of Information Management, Beijing Normal University, Beijing 100875, China*
[2]*The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518057, China*
[3]*Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China*
jinjian.jay@bnu.edu.cn, mfpji@polyu.edu.hk

Keywords: customer requirement, online review analysis, text mining

Abstract

Online product reviews contain valuable information about customer requirements (CRs). Intelligent analysis of a large volume of online CRs attracts interest from researchers in different fields. However, many research studies only concern sentiment polarity in different level and designers still need to read these reviews to absorb comprehensive CRs. In this research, online reviews are analyzed to obtain consumers' fine-grained concerns. Specifically, aspects of product features and detailed reasons of consumers are extracted from online reviews. This research starts from the identification of product features and the sentiment analysis with the help of pros and cons reviews. Next, the approach of conditional random fields is employed to detect aspects of product features and detailed reasons jointly. In addition, a co-clustering algorithm is devised to group similar aspects and reasons to provide concise descriptions about CRs. Finally, with hundreds of customer reviews of six mobiles in Amazon.com, a case study is presented to illustrate how the proposed approaches benefit product designers in the elicitation of CRs by the analysis of online opinion data.

1 Introduction

The rapid development of ebusiness makes a large volume of online reviews are generated constantly online. These reviews contain valuable customer requirements (CRs) and they may help designers understand CRs which alleviate from the time-consuming investigations. But it is generally difficult to read all reviews. If some are not covered, critical messages might be neglected. This problem interests researchers in different fields. Specially, one trendy research topic is opinion mining, whose concerns are mainly on how to extract product features, how to identify sentiment polarity in the product feature level from textual data [5, 15], etc.

Besides, some researchers try to provide a concise summarization that covers consumers' major concerns. For instance, a review summarization framework was proposed at a sentence level [24]. Opinions were captured from the expansion of seed words from the WordNet. Next, dependency relation templates were utilized to detect feature-opinion pairs. Finally, these pairs were utilized to organize sentences, which were deemed as the review summary. Another summarization approach was reported to analyze the topic structure of online reviews [22]. Important topics were extracted and aggregated from online reviews. The final summary of reviews was clustered by the topic structure. A probabilistic mixture model was initially proposed to analyze topics and sentiments in online reviews [14]. In this model, a document was considered to be generated by background words and other words, which were generated from one of many subtopics. Next, a sentiment word was utilized to describe the topic. Finally, a HMM (Hidden Markov Model) model was employed to analyze the dynamic change of the sentiments in online reviews. Jakob and Gurevych utilized tokens, POS (Part of Speech) tags, short dependency paths, word distances between opinion words and other features to define features [6]. With these features, the approach of CRFs was to detect opinion targets in online reviews. Different structures of CRFs were are found to be to utilize to identify product features and related opinion words, which were regarded as a kind of review summarization [1, 10].

Some researchers also try to analyze online opinion data for product designers. A framework was presented to aggregate CRs from online reviews for product design [2]. This framework was to infer the relative effect of product features and effect of different brands on overall customer satisfaction. A system that monitors customer opinions from textual data was built [4]. First, whether a phrase was interesting is defined on the frequency, its previous referred frequency and the level of specificity. Next, phrases near the terms of interest were extracted and they were utilized to identify which of them will emerge dramatically. Besides, how to prioritize engineering characteristics from reviews was investigated [7]. The rating values of online reviews were regarded as the overall customer satisfaction and an ordinal classification approach was proposed to prioritize engineering characteristics for designers. The helpfulness of online reviews was initially defined from the perspective of designers [13]. The helpfulness of online reviews was inferred by using a regression method. They claimed that, with domain-independent features only, it is found that there was no significant loss in terms of the helpfulness prediction. Online opinion data are also utilized to make comparisons between products. For instance, utilizing online reviews, researchers extracted product names and product features using CRFs [16]. With the proposed approaches, two applications were described to compare products. Some other approaches based on the graph propagation were proposed to compare products by using online reviews [11, 23]. To capture the rapid change of CRs, a two-stage hierarchical process was built from online reviews [9]. At the first stage, related product attributes and CRs were clustered into hyper-edges by an association rule algorithm. At the second stage, hyper-rules were applied on hyper-edges to track CRs. Similarly, online reviews were utilized to predict product design trends [18]. The sentiment polarity in the feature level was extracted from product reviews. Next, the Holt-Winters exponential smoothing method was used to model product preference trends.

Nonetheless, there exist other practical concerns from product designers. Consider the following review sentences of one mobile in Amazon.com for example.

S1 : "The battery life is a horrible."

S2 : "The on-board battery meter can be misleading."

S3 : "My only complaint here is that the battery is difficult to remove."

In this example, most opinion mining techniques are capable of recognizing that consumers are writing negative opinions regarding the battery of a mobile phone. But designers still need to consolidate all of the reviews and read these sentences to identify consumers' concerns. For example, different aspects of the battery are critiqued in S1-S3 and these details provide instructive suggestions. However, many relevant studies fail to differentiate aspects of the battery and note the detailed reasons about what makes consumers unsatisfied.

Automatically identifying different aspects of product features from online reviews helps to clarify CRs for designers further. In addition, the analysis of consumers' detailed reasons that support their arguments is also helpful for designers to comprehend CRs. Hence, the review analysis needs to be conducted at a fine-grained level to explore consumers' fine grained concerns. In particular, four types of messages should be exploited. Generally, let Q = <F, S, A, R> be a quadruple, which includes product features (F), sentiment polarity (S), aspects of product features (A) and detailed reasons (R). Specially, an aspect refers to one property of the product feature, such as battery life, screen size, etc. The detailed reasons refer to the explanations that are written by consumers to support their arguments regarding their sentiment polarity.

Take the previous three sentences for example. In the first sentence, a negative polarity is presented on the battery life and no extra details are provided. Then, the extracted quadruple will be Q = <"battery", "negative", "life", NULL>. NULL denotes that no additional reasons are provided by the consumer. In the second sentence, the consumer complaints about the battery meter and the consumer utilizes the word "misleading" to support his/her argument. Hence, the quadruple of the second sentence is Q = <"battery", "negative", "meter", "misleading">. For the third sentence, the complaint is concerning the difficulty of removing the battery. Thus, the quadruple is Q = <"battery", "negative", NULL, "remove">.

The problem to be investigated is the extraction of four types of information from online reviews. Specially, four tasks need to be conducted, including (1) to extract product features, (2) to identify sentiment polarity, (3) to recognize aspects of product features and (4) to discern detailed reasons of consumers.

2 Methods

2.1 A Framework for the Identification of Feature Aspects and Consumer Reasons

In Figure 1, a framework for the identification of product feature aspects and consumer detailed reasons from online reviews is presented. As seen from this figure, POS tags are initially obtained from online reviews. These tags are utilized for the identification of product features and the analysis of sentiment polarity. In this study, with the help of pros and cons reviews, a supervised learning approach is devised to identify product features and analyse the corresponding sentiment polarity.

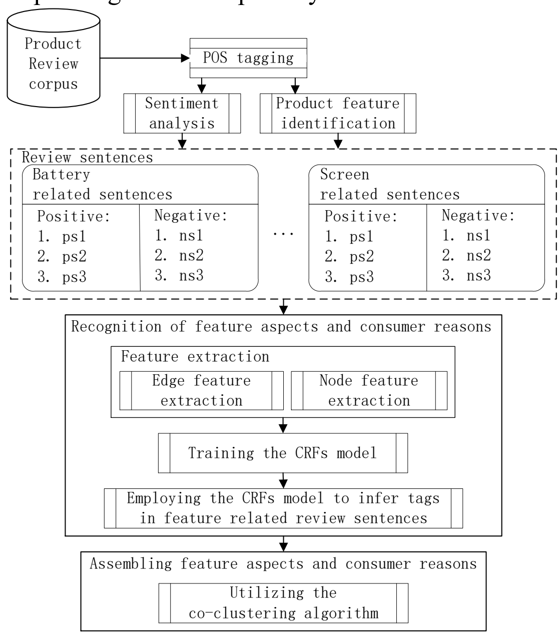


Figure 1. A Framework for the identification of product feature aspects and consumer detailed reasons from online reviews

Note that, concentrations of this study are to recognize aspects of product features and to discern detailed reasons of consumers from online reviews. Specially, a two-phase approach is suggested. In this first phase, a CRFs model was trained for the recognition of feature aspects and consumer detailed reasons from each product feature related sentences. In this CRFs based model, two categories of features were utilized. The node features characterized the word related information. The edge features were employed to depict the association between hidden tags and the association between a hidden tag and a visual word. Next, the trained model was utilized to infer a quadruple Q = <F, S, A, R> from each product feature related sentence. Technical details are explained in Section 2.3. In the second phase, a co-clustering algorithm was proposed. This algorithm was to jointly assemble recognized similar product feature aspects and consumer detailed reasons. This algorithm provides a brief summary towards consumers' concerns on different aspects of product features.

2.2 Extracting Product Features and Identifying Sentiment Polarity

In the area of opinion mining, different approaches are proposed to extract product features and identify sentiment polarity [3, 12, 21]. However, some approaches are slightly difficult for data practitioners in other fields to understand and implement. Additionally, similar to other supervised learning approaches, a large number of labeled data are required for model training, which is generally time-consuming to obtain. In this research, the extraction of product features and the identification of sentiment polarity are conducted with the help of pros and cons reviews. Similar approaches are also explained in [8, 20].

2.2.1 Extracting Product Features

In this research, pros and cons reviews are utilized to extract product features. A representative pros and cons review can be found in Epinions.com. As noted, product features are highlighted clearly in the pros and cons lists. These lists provide valuable training data and help to extract product features from reviews. Accordingly, with the POS tagging conducted on reviews, the frequently mentioned nouns or noun phrases in the pros and cons list are used to identify product features.

Additionally, different words are employed to describe the same product features. For instance, both "memory" and "storage" are utilized interchangeably to denote the same feature. With the help of WordNet, synonyms of nouns are extracted and these synonyms are clustered to identify the same product feature. To avoid imprecise expanded synonyms of words, only synonyms within two WordNet distances are considered. In addition, abbreviations are often utilized in reviews. For instance, "apps" is used to denote the "applications" of mobiles. However, these words are seldom defined in WordNet or other thesauruses. Thus, some manually defined rules are made to supplement synonyms from the WordNet expansion. Finally, product features are extracted from online reviews and clustered by using pros and cons reviews.

2.2.2 Identifying Sentiment Polarity

In this research, it is assumed that a consumer holds a neutral sentiment if one sentence has an objective opinion. Two subtasks need to be conducted for the identification of sentiment polarity. One is to know whether a subjective or objective opinion is expressed. The other is to identify whether consumers hold a positive or negative opinion.

In 2004, Pang and Lee built a publicly available subjective dataset, which included 5,000 subjective and 5,000 objective sentences [17]. Hence, for the first subtask, with the help of this dataset and denoting each sentence as a bag of words (BOWs), a binary Naive Bayes classifier can be constructed to distinguish subjective and objective sentences in online reviews.

If a sentence is predicted to be subjective, the next subtask is to judge whether it is positive or negative. Notice that besides the product features listed in pros and cons reviews, the positive and negative sentimental information about features is also provided. Then, utilizing sentences with sentiment polarity as training data, a binary classifier is built to discriminate whether a positive opinion is expressed in each sentence. To build this sentiment classifier, rather than a BOWs representation, sentimental terms are considered in review sentences. Notably, in this research, each sentence is represented by sentimental terms in the subjective lexicon provided by the MPQA project [19]. Using sentences with sentiment polarity as training data and MPQA representation, a binary classifier is built to discern the sentiment polarity of product features.

2.3 Recognizing Aspects of Product Features and Detailed Reasons of Consumers

As explained, the extraction of aspects and reasons requires identifying a sequence of words that are evaluated as properties of features or arguments to back up consumers' opinions. In this research, an approach of CRFs is utilized to tag each word in the review sentences.

2.3.1 Labeling Scheme

Three types of expressions are required to be identified from online reviews, including product features, aspects of product features and reasons of consumers. In this research, the BIO encoding for tag representation is utilized to label each sentence containing a product feature. Specially, product features are represented as "F"s and aspects of product features are "A"s. For a sequence of words that describes reasons, "RB" is to represent the beginning of one reason and "RI" is for the inside of the reason. All other words and punctuations are then labeled as "O"s. Finally, in total, five tags are utilized to denote each word in a review sentence. Following this labeling scheme, the previous three sentences in Section 1 can be denoted as,

S1: The/O battery/F life/A is/O a/O nightmare/O ./O

S2: The/O on-board/O battery/F meter/A can/O be/O misleading/O ./O

S3: My/O only/O complaint/O here/O is/O that/O the/O battery/F is/O difficult/O to/O remove/RB ./O

Another complex example is also shown as follows.

S4: Using/O features/O such/O as/O playing/RB games/RI or/O watching/RB movies/RI can/O really/O drain/RB the/O battery/F life/A ./O

2.3.2 Feature Extraction

In this subsection, node features and edge features extracted from online reviews for tag labeling are described.

All node features are listed in Table 1. Position features are to judge the position of the current token. In considerations of the current token, the previous token and the following token, six types of word features are utilized. The token category is determined according to the shape of a token. In this research, the 19 categories in LingPipe are utilized, which include whether it is a single digit, whether it contains letter only, etc. Moreover, four types of prefixes and suffixes include sequences containing one, two, three and four characters, respectively. Dependency relation features provide detailed information about the parsed dependency tree. In this research, Stanford dependencies representation is utilized, which is often employed to analyze grammatical relationships in a sentence. In this representation, a triple denotes the relationship between a pair of words, which involves the name of this relation, the governor of this relation and the dependent of this relation.

Table 1. Node features

Position features	Is the first token
	Is the last token
	Is not the first token or the last token
	The position of the current token
Word features	Current token
	Previous token
	Next token
	Lemma of the current token
	Lemma of the previous token
	Lemma of the next token
	Token category of the current token
	Token category of the previous token
	Token category of the next token
	Current POS tag
	Previous POS tag
	Next POS tag
	Prefixes of the current token
	Prefixes of the previous token
	Prefixes of the next token
	Suffixes of the current token
	Suffixes of the previous token
	Suffixes of the next token
Dependency relation features	Is a governor
	Name of relation if it is a governor
	Token of the dependent if it is a governor
	Name of relation and token of the dependent if this node is a governor
	Is a dependent
	Name of relation if it is a dependent
	Token of the governor if it is a dependent
	Name of relation and token of the governor if this node is a dependent

For edge features, they are presented in Table 2. The aim of tag features provides the tag and category information of the previous token. Similarly, dependency relation features define the tag and the name about the relation for the case in which a token is involved as a governor or a dependent of one grammatical relationship.

Table2. Edge features

Tag features	Tag of the previous token
	Token category and the tag of the previous token
Dependency Relation features	Tag of the dependent if it is a governor
	Name of relation and tag of the dependent if the node is a governor
	Tag of the governor if it is a dependent
	Name of relation and tag of the governor if the node is a dependent

2.4 A Co-clustering Algorithm for Aspects of Features and Consumer Reasons

A simple approach is to cluster consumer reasons by the WordNet distance. However, this approach cannot be applied directly. As noted, both phrases and words are utilized to describe reasons. Since the distance between phrases cannot be evaluated by WordNet, such approach fails to be applied to cluster reasons. Additionally, it cannot handle the previous use case where "few hour", "less than a day" and "drain" are employed in different sentences, although they complains about the battery life.

In this research, a co-clustering algorithm is proposed to jointly cluster reasons of consumers and aspects of product features. The intuition behind this algorithm is that similar words that describe reasons are often derived from the same aspect and, likewise, the same aspect often derives to similar words. For example, the followings are four typical sentences with labeled tags to describe the battery of one mobile.

S1: The/O on-board/O battery/F meter/A can/O be/O misleading/RB
S2: The/O battery/F indicator/A is/O deceptive/RB
S3: It/O can/O really/O KO/RB the/O battery/F life/A
S4: The/O battery/F life/A is/O not/O what/O promised/O no/O matter/O if/O you/O find/O a/O way/O to/O charge/RB it/O

In S1 and S2, it can be concluded that "meter" and "indicator" are the same aspect because the reasons of "misleading" and "deceptive" are semantically similar. Additionally, in S3 and S4, "KO" and "charge" are regarded to be similar because both words point to the battery life. Accordingly, these detailed reasons are utilized to cluster the corresponding aspects and, likewise, the aspects of product features are utilized to cluster the corresponding reasons. The details are described in Algorithm 1.

Algorithm 1: A co-clustering algorithm for aspects of features and consumer reasons

Input:
Map of aspects A_s ← <Aspect, <Document ID>>
Map of reasons R_s ← <Reason, <Document ID>>
Output:
Clusters of aspects C_a, Cluster of consumer reasons C_r
Steps:
1 DO
2 FOR any two reasons R_a and R_b in R_s
3 IF R_a and R_b are similar
4 $IDs(R_a)$ ← Extract All Document ID of R_a
5 $IDs(R_b)$ ← Extract All Document ID of R_b
6 $As(IDs(R_a))$ ← Extract All Aspects of $IDs(R_a)$
7 $As(IDs(R_b))$ ← Extract All Aspects of $IDs(R_b)$
8 A_s ← Group $As(IDs(R_a))$ and $As(IDs(R_b))$
9 END IF
10 END FOR
11 C_a ← Extract clusters of aspects from A_s
12 IF A_s changed
13 FOR any two aspects A_p and A_q in A_s
14 IF A_p and A_q are similar
15 $IDs(A_p)$ ← Extract All Document ID of A_p
16 $IDs(A_q)$ ← Extract All Document ID of A_q
17 $Rs(IDs(A_p))$ ← Extract All Reasons of $IDs(A_p)$
18 $Rs(IDs(A_q))$ ← Extract All Reasons of $IDs(A_q)$
19 R_s ← Group $Rs(IDs(A_p))$ and $Rs(IDs(A_q))$
20 END IF
21 END FOR
22 END IF
23 C_r ← Extract clusters of reasons from R_s
24 WHILE A_s or R_s changed, return to Line 1

The objective of Algorithm 1 is to cluster documents according to both aspects of product features and reasons of consumers. First, if two reasons are semantically similar, the corresponding documents are extracted (line 3 ~ 5). Aspects of product features that connect to these documents are considered to be similar and these aspects are put into the same cluster (line 6 ~ 8). Likewise, if two aspects are semantically similar, according to the referred documents, the corresponding reasons are extracted and they are grouped into the same cluster (line 12 ~ 20).

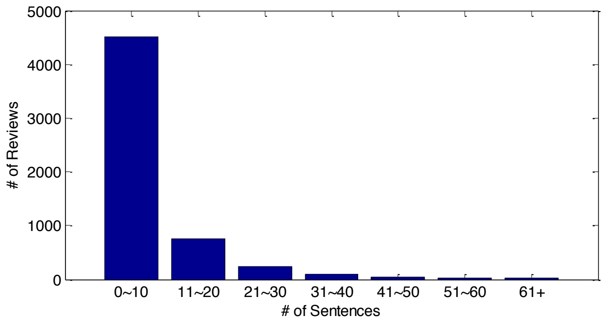
3 Case Study

3.1 Data Preparation

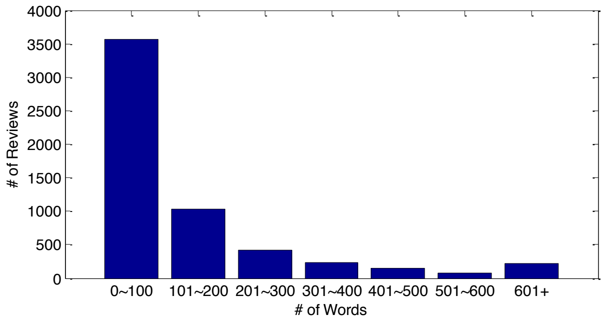
In this study, 475 pros and cons reviews of 13 smart phones were collected from Epinions.com. These pros and cons reviews are utilized as training data to extract product features and identify the sentiment polarity from reviews in a general format. Specially, 5,730 reviews of six mobile phones in Amazon.com are utilized. In consideration of data privacy, the names of these products are represented as P_1, P_2, P_3, P_4, P_5 and P_6. The number of reviews is listed in Table 3 and some statistics about these reviews are shown in Figure 2.

Table 3. The number of reviews

Name	P_1	P_2	P_3	P_4	P_5	P_6
#of reviews	1108	784	880	965	905	1088



(a) # of reviews vs. # of sentences



(a) # of reviews vs. # of words

Figure 2. Statistics of 5,719 reviews of six mobiles

On the average, there exist 142.16 words in each review. However, they are not distributed evenly, with the maximum of 3,379 words in a single review. A similar phenomenon is also found in terms of the sentence number per review, with an average 7.32 and a maximum 239.

For the labeling task, three annotators were hired to label reviews manually. The labeling scheme is introduced in Section 2.3.1. Each review was labeled by two annotators. Conflicts of review labeling were examined and determined by the third annotator.

3.2 Product Feature Extraction and Sentiment Identification

The top five frequently discussed features of 5,730 mobile reviews are shown in Table 4. As seen from this table, screen, application and battery become hot features. Generally, all six selected products are smart phones. Perhaps the first impression of a smart phone is of its screen. Indeed, a large and clear screen makes it attractive to consumers. Additionally, applications in the phone and its operating system are other critical factors that consumers use to make purchasing decisions. For instance, comparisons of applications in different operating system, such as Android or Windows, affect consumers' decisions on the selection of brands and product models. Moreover, if one mobile's battery bother consumers to charge regularly, it can be expected that it will result in negative comments.

Table 4. Top five frequently discussed features

Top referred product features	% of reviews referred features
cover screen screens	24.41%
app application applications apps	18.38%
batteries battery	17.96%
android androids os symbian window windows	17.68%
internet net network networks web wi-fi wifi	15.70%

Accordingly, the battery and the screen are chosen as exemplary product features. The objective is to exemplify the extraction of aspects and the identification of reasons that lead consumers to provide a negative sentiment. In Table 5, some statistics are shown.

Table 5. Statistics about feature related reviews

Product	Feature	# of reviews referred this feature	% of reviews referred this feature	% of negative towards the feature
P_1	Screen	324	29.24%	51.23%
	Battery	105	9.48%	50.48%
P_2	Screen	218	27.81%	44.95%
	Battery	206	26.27%	56.31%
P_3	Screen	211	23.98%	50.24%
	Battery	172	19.55%	40.12%
P_4	Screen	188	19.71%	42.55%
	Battery	149	15.62%	48.99%
P_5	Screen	165	18.23%	45.45%
	Battery	154	17.02%	50.65%
P_6	Screen	290	26.65%	24.48%
	Battery	241	22.15%	33.61%

More than 35% consumers prefer to talk about either of these two product features. However, the sentiment polarity of consumers towards two features are not the same. The percentage of negative reviews holds an indispensable share. To investigate what make consumers unsatisfied regarding a specific product in feature level, designers still need to read these negative opinions and explore reasons for understanding CRs.

In Table 6, several exemplary sentences for the battery and the screen are sampled from the negative reviews of six products. Consider the screen, for instance. Different aspects are mentioned, including response time, automatic locking, ease of scratched, size, resolution, and rotation. It further confirms that different aspects of product features lead to negative opinions and the detailed reasons provide several instructive suggestions to designers.

Table 6. Some exemplary negative reviews

Screen	1. The touch screen is so slow and so hard to use. 2. The biggest problem with this phone is that the screen automatically locks after you dial a phone number to call. 3. Plastic touch screen can be easily scratched , use caution in pocket with keys or coins. 4. The screen is so small , that even with my reading glasses , I have a horrible time reading anything. 5. Resolution of the screen again is not as good as the iphone. 6. Sometimes when I flip the phone , the screen will not rotate.
Battery	1. Biggest complaint about this phone is it 's ridiculously terrible battery life. 2. They thought the battery was too thick. 3. Some of the more advanced features are particularly battery draining. 4. Had problems with the battery after a solid year of use, which I had to replace but it was n't expensive (16 bucks) . 5. In my honest opinion , it looked like a cheap battery. 6. I am also extremely disappointed in the fact that the battery is embeded and therefor not replaceable.

3.3 Exploration of Aspects of Product Features and Reasons of Consumers

To evaluate the performance of the proposed method, in this case study, review sentences that are talking about the battery and the screen of P_1, P_2 and P_3 are examined.

3.3.1 Evaluation Metrics

Three widely utilized classification evaluation metrics are employed, including precision, recall and F_1. Precision and recall are generally defined according to the measure of relevance. Precision is the fraction of retrieved instances that are relevant. Precision is denoted as

$$\text{Precision} = \frac{|\{\text{relevant documents}\}| \cap |\{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Recall is the fraction of relevant instances that are retrieved,

$$\text{Recall} = \frac{|\{\text{relevant documents}\}| \cap |\{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

F_1 is the harmonic mean of precision and recall.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

3.3.2 Evaluation Results

The following results are reported on the average of cross evaluations over products. Specially, it takes reviews of two products as training data and reviews of the third product as testing data.

The performance of the proposed CRFs based approach is listed in Table 7 and Table 8, respectively. As seen from the two tables, words that refer to product features are recognized accurately in both datasets.

Table 7. Aspects and reasons identification over reviews about the battery

Battery	Feature	Aspect	Reason
Precision	1.000	0.936	0.787
Recall	0.998	0.854	0.453
F_1	0.999	0.893	0.575

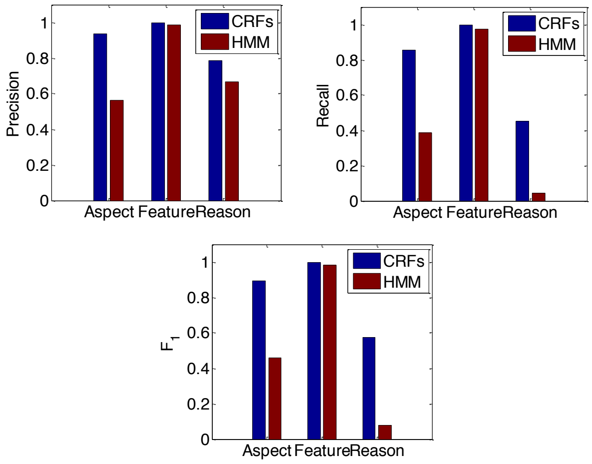
Table 8. Aspects and reasons identification over reviews about the screen

Screen	Feature	Aspect	Reason
Precision	0.990	0.580	0.569
Recall	0.985	0.444	0.306
F_1	0.988	0.503	0.398

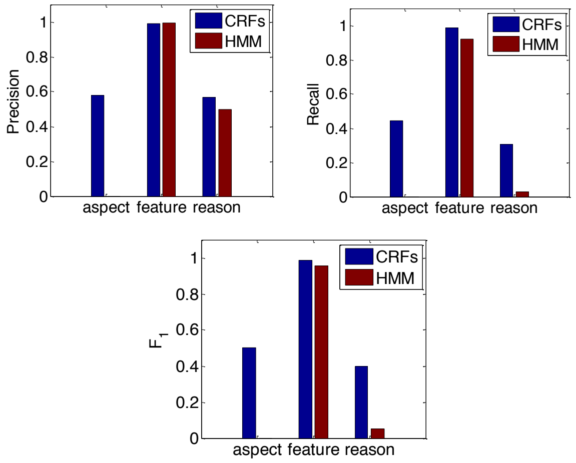
In terms of the identification of aspects, a relatively higher performance is achieved in the battery dataset. As noted, words that describe aspects of battery tend to be limited. For instance, these phrases include "battery life", "battery charging", "battery drain", "battery alert" and "battery maintenance". However, for phrases that describe a screen, they tend to be fuzzy and include "screen space", "screen lock", "screen size", and "screen display". In addition, the complexity of phrase structures that are utilized to describe aspects of two features is also different. For example, in "life of battery", it is relatively easy to infer that the word "life" tends to be one aspect of the battery. Similarly, "size" in "the size of screen" can be deduced as one aspect of the screen. However, in another case, "corner" in "the right corner of screen" is not an aspect, although this phrase has a similar structure. The diversity of words and the complexity of structures make it somewhat difficult to detect aspects accurately.

For the identification of the detailed reasons, compared with the recall, a higher precision is obtained in both datasets. A wild range of words are utilized to describe the detailed reasons, which make them tend to be quite literally dissimilar with each other, although the same aspect of product features is pointed to. Additionally, occasionally, complaints of consumers are twisted with different product features. For instance, one consumer complaint said that the "QUERTY keyboard is too small for efficient usability, Camera is difficult can't see the icons on the screen". Generally, these complaints describe that the icon formatting on the screen makes it difficult for consumers to use the phone, although the screen of this mobile is mentioned. Another case is that some consumers explain the whole scenario about what makes him/her unsatisfied, such as "from having the phone in your pocket, buttons are pressed and it turns the screen on, which uses some of your battery". This example is describing the sensitivity of the screen, though no specific words are employed to describe the aspects of the product features directly. All of these complex structures make it difficult to identify reasons at a high recall. It leaves some space to develop sophisticated models to enhance the overall performance.

As noted in Section 2, a sequential label inference problem for textual data is modeled. Notably, the HMM is also a frequently utilized approach to infer hidden states for sequential tagging. Accordingly, to make comparisons with the proposed CRFs based approach, an HMM based approach was conducted. The performance comparison is reported in Figure 3.



(a) Performance comparison on the battery dataset



(b) Performance comparison on the screen dataset

Figure 3. Comparison of the CRFs based approach and the HMM based approach

As seen from this figure, the proposed CRFs based approach outperformed the HMM based approach in both two datasets. High precision and recall were gained by both approaches for the recognition of product features from review sentences. However, for the recognition of aspects of product features and detailed reasons of consumes, the CRFs based approach were seen to perform much better than the HMM based approach. Especially, in the screen dataset, the HMM based approach nearly failed to recognize aspects of product features. As explained, only a few words were employed to denote a specific product feature within these two feature related datasets, which makes CRFs and HMM successfully label the correct words or phrases. However, fuzzy phrases and words tends to be utilized to describe aspects of product features and detailed reasons of consumers. Additionally, in this study, five tags were utilized to label different perspectives of consumers' concerns and only the transition probability between two successive tags are reckoned in the HMM based approach. Sufficient correlated information in review corpus is not captured. It might be the major reasons that leads to a relative poor performance on the recognition of aspects of product features and detailed reasons of consumes. Besides, more diversified phrases tends to be utilized in the screen dataset, which aggravate the poor performance.

In the previous sections, the objective of clustering identified aspects of product features and detailed reasons of consumers is highlighted to provide a concise description about CRs. In the following, battery and screen related review sentences of P_1 and P_2 are utilized as training data and that of P_3 are utilized as testing data. With the proposed co-clustering algorithm, aspects of battery and screen as well as detailed reasons of consumers are listed in Table 9 and in Table 10, respectively.

Table 9. Identified aspects of battery and the detailed reasons of consumers

Aspects	{indicator}, {model}, {meter}, {maintenance}, {LIFE; life}, {power}, {load}, {saver}, {design}
Reasons	{GPS}, {wifi; internet}, {texting}, {application}, {call}, {charge; charging; charged; drained; drain; last; recharging; recharge; charger; discharge; KO}, {game}, {# hour; #-# hour; few hour; a couple of hour}, {email}, {direction}, {less than a day; a night; long more than # day; a day; # day; #.# day; # - # day; day}, {# pm}, {dead; died; run; bother; change; replace; get; replaceable}, {save}, {# year}, {widget}, {remove; removable}

Table 10. Identified aspects of screen and the detailed reasons of consumers

Aspects	{brightness}, {keyboard}, {suck}, {protector; saver}, {side}, {lock}, {quality}, {prorifery}, {resolution}, {size}
Reasons	{resolution/brightness}, {see; Playing; freezed; used; flip; dying; rotate; Locked; read; created; turned; show; freeze; unlock}, {saver}, {responsiveness; response; unresponsive; scratch; responding; respond to touch; stuck}, {plastic}, {located}, {bright}, {impressive}, {bigger; small; big; large; larger}, {sluggish}, {resistive}, {cracked}, {sensitivity; sensitive}, {capacitive}, {defective}, {black}, {advertised}, {corrupted}, {blank; off}, {clean}

A number of detailed reasons that complain about the battery life are presented as clusters of words in Table 10. In particular, all the digits were replaced by the symbol "#" before the co-clustering algorithm was applied. These clusters are employed to describe CRs, such as "charge", "KO", "hour" and "day". Additionally, other reasons are frequently mentioned by consumers, such as "GPS", "internet", "texting", "application", "call", "game" and "email". These reviews actually point to the battery consumption by some specific functions of P_3. It suggests that battery designers of P_3 should check why these functions drain the battery quickly or provide a solution to prevent the battery being consumed too fast. There also exist consumer complains about the difficulty of replacing or removing the battery. Thus, a user-friendly design of battery replacement is recommended. Compared with reviews that point to the battery of a mobile, reviews that complain about the screen are diverse. A number of consumers are dissatisfied with the size of the screen by describing it as, for example, "small", "big" and "larger". Certain consumers are disappointed with the response of the screen with descriptive words, such as "response" and "stuck". Intuitively, consumers refer to the screen's "sensitivity". However, generally, the word "sensitivity" and the word "response" are not literally defined to be similar to each other. Accordingly, as shown in Table 10, they are inferred as two clusters. However, in a specific domain, like one where reviews refer to the screen of mobile phones, these words should be grouped in the same cluster. Hence, in the future, domain-specific similarities between words need to be considered in sophisticated algorithms.

4 Conclusions

Online opinions are generated from time to time, which contain valuable CRs about products. Effectively understanding CRs at a fine-grain level based on online opinions exert an influential aspect on the improvement of products in market-driven design.

The objective of this research is to extract various aspects of product features and investigate detailed reasons regarding what make consumers unsatisfied with products. Particularly, product features and correlated sentiment polarity are identified with the help of pros and cons reviews. Next, an approach based on CRFs is used to pinpoint aspects of features, as well as reasons of consumers from online reviews. Furthermore, a co-clustering algorithm is proposed to group jointly both aspects of features and reasons of consumers to provide a concise description regarding CRs for product designers. In addition to the extraction of sentiment polarity in product feature level, this research enables designers to obtain more insightful suggestions from online reviews and facilitates them to absorb CRs from big opinion data efficiently.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. NSFC 71271185/G011201), the Fundamental Research Funds for the Central Universities (No. SKZZX2013091) and the Research Project of The Hong Kong Polytechnic University (No. G-YBE3).

References

[1] Chen, L. and Qi, L. Social opinion mining for supporting buyers complex decision making: exploratory user study and algorithm comparison. Social Network Analysis and Mining, 1(4):301-320, 2011

[2] Decker, R. and Trusov, M. Estimating aggregate consumer preferences from online product reviews. International Journal of Research in Marketing, 27(4):293-307, 2010

[3] Ding, X., Liu, B., and Yu, P. S. A holistic lexicon-based approach to opinion mining. In WSDM'08, 2008

[4] Goorha, S. and Ungar, L. Discovery of significant emerging trends. In KDD'10, pages 57-64, 2010

[5] Hu, M. and Liu, B. Mining and summarizing customer reviews. In KDD'04, pages 168-177, 2004

[6] Jakob, N. and Gurevych, I. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In EMNLP'10, pages 1035-1045, 2010

[7] Jin, J., Ji, P. and Liu, Y. Product characteristic weighting for designer from online reviews: an ordinal classification approach. In EDBT/ICDT'12 Workshop: BEWEB, Berlin, Germany, 2012

[8] Kim, S.-M. and Hovy, E. Automatic identification of pro and con reasons in online reviews. In COLING'06, pages 483-490, 2006

[9] Lee, T. Needs-based analysis of online customer reviews. In ICEC'07, pages 311--318, 2007.

[10] Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S., and Yu, H. Structure-aware review mining and summarization. In COLING'10, pages 653-661, 2010

[11] Li, S., Zha, Z.-J., Ming, Z., Wang, M., Chua, T.-S., Guo, J., and Xu, W. Product comparison using comparative relations. In SIGIR'11, pages 1151-1152, 2011

[12] Lin, C. and He, Y. Joint sentiment/topic model for sentiment analysis. In CIKM'09, pages 375-384, 2009

[13] Liu, Y., Jin, J., Ji, P., Harding, J. A. and Fung, R. Y. K.. Identifying Helpful Online Reviews: A Product Designer's Perspective. Computer-Aided Design, 45(2):180-194, 2013

[14] Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. Topic sentiment mixture: modeling facets and opinions in weblogs. In WWW'07, pages 171-180, 2007

[15] Moghaddam, S. and Ester, M. Aspect-based opinion mining from product reviews. In SIGIR'12, pages 1184-1184, 2012

[16] Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. Mine your own business: Market-structure surveillance through text mining. Marketing Science, 31(3):521-543, 2012

[17] Pang, B. and Lee, L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In ACL'04, 2004

[18] Tucker, C. and Kim, H. M. Predicting emerging product design trend by mining publicity available customer review data. In ICED'11, pages 43-52, 2011

[19] Wilson, T., Wiebe, J., and Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In EMNLP'05, pages 347-354, 2005

[20] Yu, J., Zha, Z.-J., Wang, M., and Chua, T.-S. Aspect ranking: identifying important product aspects from online consumer reviews. In ACL'11, pages 1496-1505, 2011

[21] Zhai, Z., Liu, B., Xu, H., and Jia, P. Grouping product features using semi-supervised learning with soft-constraints. In COLING'10, pages 1272-1280, 2010
[22] Zhan, J., Loh, H. T. and Liu, Y. Gather customer concerns from online product reviews - a text summarization approach. Expert Systems With Applications, 36(2 Part 1):2107--2115, 2009
[23] Zhang, Z., Guo, C., and Goes, P. Product comparison networks for competitive analysis of online word-of-mouth. ACM Transactions on Management Information Systems, 3(4):1-20, 2013
[24] Zhuang, L., Jing, F., and Zhu, X. Y. Movie review mining and summarization. In CIKM'06, pages 43-50, 2006