Identifying Product Defects from User Complaints: A Probabilistic Defect Model

Xuan Zhang[1], Zhilei Qiao[2], Lijie Tang[1], Patrick (Weiguo) Fan[1,3], Edward Fox[1], Alan (Gang) Wang[2]

1, Department of Computer Science, Virginia Tech

2, Department of Business Information Technology, Virginia Tech

3, Department of Accounting and Information Systems, Virginia Tech

{xuancs, qzhilei, lijiet, wfan, fox, alanwang}@vt.edu

2 March 2016

Abstract

The recent surge in using social media has created a massive amount of unstructured textual complaints about products and services. However, discovering and quantifying potential product defects from large amounts of unstructured text is a nontrivial task. In this paper, we develop a probabilistic defect model (PDM) that identifies the most critical product issues and corresponding product attributes, simultaneously. We facilitate domain-oriented key attributes (e.g., product model, year of production, defective components, symptoms, etc.) of a product to identify and acquire integral information of defect. We conduct comprehensive evaluations including quantitative evaluations and qualitative evaluations to ensure the quality of discovered information. Experimental results demonstrate that our proposed model outperforms existing unsupervised method (K-Means Clustering), and could find more valuable information. Our research has significant managerial implications for mangers, manufacturers, and policy makers.

[Category: Data and Text Mining]

Introduction

With the fast development of online communities and other types of social media, users can freely provide their feedback on defects of products and services (Abrahams et al. 2013; Chen et al. 2009; Yan et al. 2015). This feedback is useful to other consumers for decision-making, and to firm executives for improving their product or service quality. For example, Chen and Xie (2008) propose that user reviews work as a form of “sales assistance” to customers because they can use reviews to know the real quality of products, and reduce the uncertainty of risks and transaction costs. Abrahams et al. (2012) demonstrate that vehicle defect discovery from social media could improve automotive quality management. As user generated content in social media surges explosively and spreads virally at an unprecedented speed, many companies seek to transform business opportunities by uncovering hidden value from the content (Fan and Gordon 2014; Luo et al. 2013; Yu et al. 2013). User generated content is mostly a type of unstructured data that is not easy to understand and manage. Therefore, effectively and efficiently extracting valuable knowledge from unstructured data has wide utility. As a result, there is a new wave of commercial data analytics companies, such as CarComplaints.com, Topsy.com, and GNIP.

However, discovering and quantifying potential product defects from large amounts of unstructured text is a nontrivial task, including for academic researchers. They have struggled with this difficult problem, since it is not amenable to human text perusal, even for moderately sized textual content. For example, Abrahams et al. (2012) manually tagged 900 discussion threads from 1,316,881 in the HondaTech.com discussion forum. This sample is far less than 0.1% of the total number of discussion threads.

Therefore, it is desirable to apply automatic content analysis to this problem. Some researchers have started to explore this research area. In order to recognize product features and customers’ opinions in customer reviews, Li et al. (2010) proposed a CRF-based review summarization approach. These methods rely on a large volume of labeled data. In terms of unsupervised learning, Yang and Cardie (2013) presented a joint inference model, and Liu et al. (2014) introduced a graph co-ranking method to identify the opinion targets and opinion words. These solutions work well when extracting product features and customers' assessments from individual product reviews, but they have difficulty in identifying common opinions from massive data. Nevertheless, it is still not able to acquire integral defect information such as: who has the defect, when the defect occurs, and what are the symptoms of the defect.

To bridge this gap, we put forward a novel method to automatically generate a defect summary. We first develop a multi-aspect topic model that utilizes the multi-aspect information of a defect, and summarizes massive complaints. For example, a defect of an automobile usually contains facts about defective components, vehicle model year, and symptom. The summary in multi-aspect will be more readable and representative for the defect. Simultaneously, we apply TF-IDF and Part-Of-Speech (POS) tagging to remove noise from complaint text. We showcase the method by analyzing user complaints of National Highway Traffic Safety Administration (NHTSA). Distinguished from previous work, we not only find the reprehensive sentence (P. Li et al. 2010) for a defect, but also provide more concrete information about a defect across different manufacturers. Our empirical results indicate that the proposed method can function better than existing unsupervised methods for analyzing unstructured user-generated content in social media.

This study makes the following contributions for defect discovery from social media. First, we introduce a novel unsupervised machine learning method, called probabilistic defect model (combined with the expectation maximization, EM, algorithm), for identifying product defects and relevant details from unstructured textual user complaints. To the best of our knowledge, this is the first work that integrates unsupervised learning methods into the quality management field. Second, we conduct a comprehensive evaluation of our proposed model according to quantitative and qualitative evaluation methods. Experimental results show that our proposed model outperforms the competing method and discovers more meaningful product defects. This model facilitates navigation of large amounts of textual data by our target users, including firm executives, product managers, or academic researchers. Third, the proposed probabilistic defect model (with the EM algorithm) extends the automated defect discovery literature as well. It provides a comprehensive framework for incorporating contextual information (syntactic features and domain background) to uncover more meaningful topics. Last but not least, our study on defect identification from user complaints also contributes to the quality management literature. Responding to customer's complaints for product quality, companies can efficiently limit the spread of defective products, as well as improve the quality of customer engagement.

The rest of the paper is organized as follows. In Section 2, we discuss relevant prior literature. In Section 3, we present our theoretical model, and then we show the experiment and evaluation result in Section 4. Finally, we conclude the paper in Section 5, including a few managerial implications of our research and providing directions for future work.

Related Work

There is relevant work in two major research areas: (1) automatic content analysis and (2) effects of product defect discovery.

Automatic Content Analysis

Our study fits the well-known research area of automatic content analysis, aiming to extract valuable knowledge from massive unstructured text and textual information. Bao and Datta (2014) suggest that there is an increasing trend in the use of automated content (or text) analysis in social science research studies. They contend that automatic content analysis is a class of quantitative methodologies in the social science field. The analysis methodology is well developed based on various techniques, such as natural language processing, text mining, information retrieval, machine learning, data mining, etc. Despite still being in its developing stage, automated content analysis has been utilized in many research areas, including accounting, information disclosure, economics, and finance.

The most common use of automated content analysis is to summarize texts. After summarizing, user generated content can be easily quantified by using the count of defects in terms of product features (e.g., product model, product component, and defect type). For instance, to investigate the effects of product defects, researchers are interested in what types of defects are disclosed in user feedback, and then predicting the potential recall risk based on the defect disclosure distribution. The automated content analysis method can extract meaningful knowledge and quantify defect information from user reviews.

Manual summarization is very labor intensive and resource (effort) consuming. Even if we set up a clear protocol for tagging the textual data, taggers are still required to read each individual review from thousands of reviews, or even more. Automated content analysis could mitigate the cost of manual summarization, by reducing the amount of human effort. There are two main kinds of methods for automated content analysis: (1) supervised learning, and (2) unsupervised learning.

Supervised Learning Method

Supervised learning methods provide another method for summarizing documents in predefined categories. The procedure is that (1) we find experts and ask them to categorize a set of documents manually, (2) then we train a supervised model that can automatically assign categories to documents based on the trained categorical results (training set). Supervised learning methods offer several advantages over traditional dictionary methods (Grimmer and Stewart 2013). First, it is obviously context specific and therefore avoids problems when predefined dictionaries are used outside specific corpora. Specifically, scholars have to develop coding rules for the variables (categories) of interest so that coders can have a clear definition about the measurement of variables. Second, they are easy to implement with existing statistical models. Third, evaluation procedures for supervised learning are well justified and convincing.

Thanks to these advantages, supervised learning methods have been used in many research areas. In healthcare, Dai et al. (2015) design a supervised learning method to identify patients with heart disease through electronic health records. In the automotive domain, Abrahams et al. (2015) find that some key terms, product features, and semantic factors can help identify product defects, but stylistic, social, and sentiment features cannot. Still, the important assumption of supervised machine learning methods is to have a set of predefined categories (product defect types). Thus, it is not a flexible method from the automatic perspective.

Unsupervised Learning Method

Compared with supervised machine learning methods, unsupervised learning methods don’t need the assumption of predefined categories. They are a type of method that can learn major categories from different features of text without explicitly presenting categories of interest. Usually, unsupervised learning methods are also called “clustering” methods. They categorize documents based on the estimated results. Unsupervised learning methods can help identify valuable information that is perhaps understudied or previously unknown.

The challenging problem of unsupervised methods, as indicated by (Bao and Datta 2014), is that an objective function that works for multiple applications is hard to define. This is difficult to achieve because human beings are typically focused on optimal “useful” ideas under some specific contextual scenarios. Grimmer and Stewart (2013) suggest that there are two approaches to handle this problem. One approach is to allow scholars to search a big dataset to generate potential interesting and useful category labels. The other approach is to develop statistical models incorporating context-specific structures and domain-oriented knowledge. This approach needs additional information, and the variation of models (Grimmer and King 2011), but leads to remarkable clustering results. For example, Li et al. (2005) proposed a probabilistic model to model the key entities of events. In addition, Li et al. (2011) extend the Latent Dirichlet Allocation (LDA) topic model by including event time, location, and other critical or important information for better measuring the event. Distinct from the original LDA model (Blei et al. 2003), the extended model assumes that all documents of an event share the same topic mixture.

The application of unsupervised learning methods to analyze text in social science is still in its infancy. In the information system literature, Aral and Walker (2011) use unsupervised topic models to cluster the content of recommendation articles. However, they did not include any context-oriented information and use the standard LDA model to solve the challenging problem of unsupervised learning methods. In the corporate risk disclosure literature, Bao and Datta (2014) extend the standard LDA model and involve sentence structure features to discover and quantify risk types from textual risk disclosures. Experimental results show the proposed model outperforms all other competing methods, including supervised learning methods and the standard LDA model, and aids discovery of meaningful risk types.

Despite the applications of unsupervised learning methods in some research areas, we did not find prior studies that attempt to analyze user reviews for product defect discovery. Based on the prior supervised study by Abrahams et al. (2015), we report our work on product defects discovery in the data after incorporating contextual information to estimate topics. The proposed probabilistic defect model overcomes the problems of unsupervised clustering by using many domain-specific attributes that contribute to defect identification.

Effects of Product Defect Discovery

This study is also related to research on the effects of product defect discovery. As described by Abrahams et al. (2015), the business value of automated product defect discovery is associated with product competitive advantage, and then, product success and commercial success. Here, we particularly focus on the effects of product defect discovery, from user reviews, on customer relationship management and defect management.

While firm managers have begun to pay attention to customer relationship management through social media, such as user reviews, very few studies focus on product defect management through analysis of unstructured text. More importantly, the defects reported in reviews are relatively credible and so, in public communication channels, the word-of-mouth effects can quickly crash the product market. In addition, with an exponential expansion in the number of reviews, keeping the same pace to respond to individual user messages is hardly possible. Fortunately, an automated product defect discovery technique can reliably distinguish reviews that describe defects, and then managers can ensure quicker response to customer feedback. Consequently, fewer defective products will reach customers' hands; accordingly, firms can save costs for new products.

While firms are extrinsically motivated to design an automated product defect technique due to customer relationship management, firms also are intrinsically motivated to develop such a technique for defect management. Product quality is an important aspect of product competitive advantage. Product defects are very costly to companies in some industries. Thus, in the motor vehicle industry, if firms find defective units, they are mandated to report to the NHTSA and take timely rectification actions. General Motors recently reported that the cost of repairing millions of vehicles reached $1.3 billion[1]. In this case, the cost of the defect is huge, especially for a large volume of sales. Effective automated defect management techniques can help firms discover defects early and reduce the number of defective products, and thus reduce financial loss.

Methodology

Defect Definition

In order to extract key defects from datasets, we define the **4 key entities** of defects, such as vehicle model and year, component, symptom words (the smoke words in Abrahams et al., 2012), and incident date. Symptom words are a set of words that are proved to be able to describe the symptoms of a vehicle defect.

Defect Entities*: {Model- Year, Component, Symptom, Date}*

Among them, the first 4 types of entities of defect are composed by words, which usually follow the **multinomial distribution** across all the complaint documents (Li et al, 2005). In contrast, the incident date entity follows the **Gaussian Mixture Model distribution**.

Defect Identification System

Shown in Figure 1, a system is developed to identify the most critical vehicle defects from complaint records. First, the complaint data is saved in a database, which can be accessed by the analysis modules. After that,

[1] http://money.cnn.com/2014/04/24/news/companies/gm-earnings-recall/

a text pre-processing module does some cleaning of the complaint text, such as text normalization, tokenization, stop-word removal, etc. Key entities of complaints, such as vehicle model, production year, component, symptoms, and date, are extracted by an entity recognition module. Then, the parameters of the defect and entity distributions are estimated by the probabilistic inference module. Next, the estimated probabilities are used to rank the defects and their entities by the defect ranking module. The list of defects sorted by probability is shown to users.

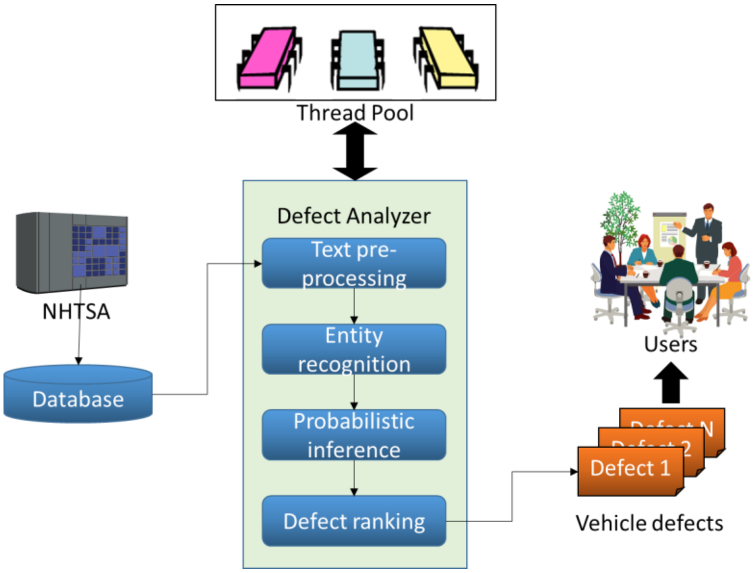

Figure 1. Product Defect Identification System

Probabilistic Defect Model

If we assume a complaint record in the "complaint" dataset is generated from a distribution of "defects", we can model their relationship by a generative probabilistic model, where "complaints" are observations and "defects" are latent variables. Complaint is defined in the same way as defect: {*Model-Year, Component, Symptom, Date*}. The first 3 entities of a complaint are modeled as vectors with their own term space. For example, given a complaint x_i, the *Symptom vector* is considered to be a list, <*symptom word 1, …, symptom word N*>, and each component *c* of the vector is the occurrence count of symptom word *c* in complaint x_i. These entities of complaints (observations) can be queried from government databases, or extracted from social media, by technologies such as regular expression matching or vehicle component isolation. Because multiple defects may be complained about at the same time, the Gaussian Mixture Model (GMM) is chosen to model the Date. We assume the 4 entities of a complaint are independent. Therefore, we have:

COMPLAINT GENERATION PROCESS

1. Choose a defect $d_j \sim Multinomial(\theta^j)$.
2. Generate a complaint $x_i \sim p(x_i|d_j)$. For each entity of it, according to the type of each entity:
   - Choose a Model-Year word: $Model_{ip} \sim Multinomial(\theta_p^j)$.
   - Choose a Component word: $Component_{iq} \sim Multinomial(\theta_q^j)$
   - Choose a Symptom word: $SmokeWord_{ir} \sim Multinomial(\theta_r^j)$
   - Draw a Date: $Date_i {\sim} N(\mu^j, \sigma^j)$

$$p(complaint) = p(model - year) * p(component) * p(symptom) * p(date) \qquad (1)$$

The generation process of a complaint record is shown by the following algorithm.

Here the *j-th* defect is represented by d_j, while the *i-th* complaint is marked as x_i; the vector θ^j represents the priors of defects; θ_p^j, θ_q^j, and θ_r^j are parameters of conditional multinomial distributions given defect d_j; μ^j and σ^j are parameters of the conditional Gaussian distribution given defect d_j.

The graphical diagram for this generative model is shown in Figure 2. Here *M*, *C*, and *S* mean the number of model-year, component, and symptom words in a complaint. The various types of words in the complaint is generated from the corresponding multinomial distribution of defects. The Date of each complaint is generated from the Gaussian distributions of the defects. *K* is the number of defects, and *D* is the number of complaints.

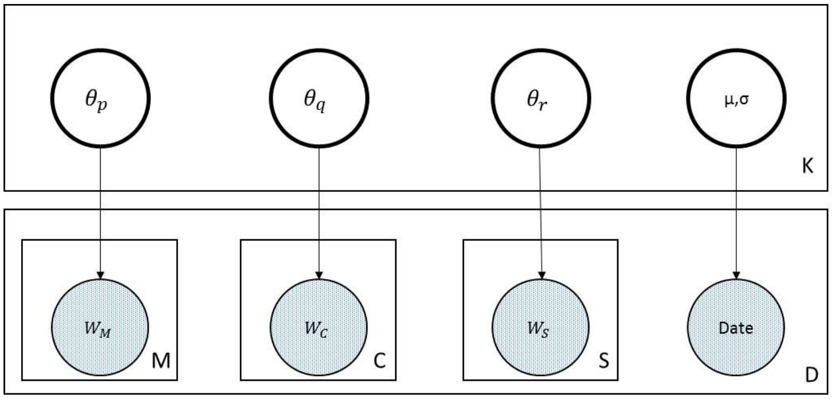

Figure 2. Graphical Diagram

Model Parameter Estimation

The model parameters can be estimated by the Maximum Likelihood method. The log-likelihood of the joint distribution of complaint is:

$$l(X;\theta) \triangleq \log\big(p(X|\theta)\big) = \log(\prod_{i=1}^{M} p(x_i|\theta)) = \sum_{i=1}^{M} log(\sum_{j=1}^{k} p(d_j)p\big(x_i\big|d_j,\theta\big)) \qquad (2)$$

In (2) *X* represents the collection of complaint records; *M* and *k* are the number of complaint records and the number of defects, respectively. These parameters can be estimated by the **EM algorithm**.

In the **E-step**, the posteriors $p\big(d_j\big|x_i\big)$ are calculated according to equations (3) (4) (5) (6):

$$p(d_j|x_i)^{(t+1)} = \frac{p(d_j)^{(t)}p(x_i|d_j)^{(t)}}{p(x_i)^{(t)}} = \frac{p(d_j)^{(t)}p(x_i|d_j)^{(t)}}{\sum_{k=1}^{K} p(d_k)^{(t)}p(x_i|d_k)^{(t)}} \qquad (3)$$

Here *(t)* means the *t-th* iteration. We need $\mathrm{p}\big(\mathrm{x_i}\big|\mathrm{d_j}\big)$ to calculate (3). According to (1), we have:

$$p\big(x_i\big|d_j\big) = p\big(model - year_i\big|d_j\big) * p\big(component_i\big|d_j\big) * p\big(symptom_i\big|d_j\big) * p(date_i|d_j) \qquad (4)$$

(i) In the case of entities following the multinomial model (e.g., *Symptom word*):

$$\boldsymbol{p\big(symptom\ word_i\big|d_j\big) = \prod_{w}^{V} p(w|d_j)^{C_w}} \qquad (5)$$

Here, *V* is the *Symptom word* set in complaint *i*; each word $w\epsilon V$ appears C_w times in the complaint *i*. The probability $p(w|d_j)$ can be initialized with a random value.

(ii) In the case of *Date* entity which follows GMM:

$$\boldsymbol{p\big(date_i\big|d_j\big) = N(date_i|\mu_j,\sigma_j)} \qquad (6)$$

Here, μ_j is the mean of the *j-th* Gaussian distribution, while σ_j is the standard deviation.

In the **M-step**, the parameters of the multinomial distributions and the GMM distribution are estimated and updated, by maximizing the $\log\big(p(X|\theta)\big)$ in (2).

(i) In the case of entities following the multinomial model (e.g., *Symptom word*):

$$p(w|d_j)^{(t+1)} = \frac{1+\sum_{i=1}^{M}(p(d_j|x_i)^{(t+1)} * tf(i,w))}{N+\sum_{i=1}^{M}(p(d_j|x_i)^{(t+1)} * \sum_{s=1}^{N} tf(i,s))} \tag{7}$$

Here *tf(i, w)* is the number of times word *w* occurs in complaint x_i, *N* is the size of that kind of vocabulary, and M is the number of complaints. **Laplace smoothing** is applied to avoid zero probabilities for infrequent words.

(ii) In the case of the *Date* entity which follows GMM:

$$\boldsymbol{\mu_j}^{(t+1)} = \frac{\sum_{i=1}^{M}(\boldsymbol{p(d_j|x_i)}^{(t+1)} * \boldsymbol{date_i})}{\sum_{i=1}^{M} \boldsymbol{p(d_j|x_i)}^{(t+1)}} \tag{8}$$

$$\boldsymbol{\sigma_j}^{2(t+1)} = \frac{\sum_{i=1}^{M}(\boldsymbol{p(d_j|x_i)}^{(t+1)} * (\boldsymbol{date_i} - \boldsymbol{\mu_j}^{(t+1)})^2)}{\sum_{i=1}^{M} \boldsymbol{p(d_j|x_i)}^{(t+1)}} \tag{9}$$

The mixture proportions (priors) are updated as in the following:

$$\boldsymbol{p(d_j)}^{(t)} = \frac{\sum_{i=1}^{M} \boldsymbol{p(d_j|x_i)}^{(t+1)}}{\boldsymbol{M}} \tag{10}$$

Once we have estimated the parameters of the probabilistic model of complaints and defects by EM, we can use the models to produce top defects, and get the key entities of each defect.

$$\boldsymbol{W = \underset{w}{argmax}(p(Model - Year_w|d_j))} \tag{11}$$

For example, for defect *j*, the most possible *Model-Year* word can be calculated by (11) and (7). The most probable words of *Component* and *Symptom* entities can be figured out in a similar way.

$$\boldsymbol{y_i = \underset{j}{argmax}(p(d_j|x_i))} \tag{12}$$

Another way to find the word most relevant to a defect is to find the complaint most relevant to a defect first. The most relevant complaint *i* given defect *j* can be figured out by (12). Here y_i is the label of complaint x_i. The complaint x_i with the maximum $p(d_j|x_i)$ among complaints assigned to the *j-th* defect is a good representative of the defect *j*.

In this way, we can get all the entities of a defect (*Model-Year, Component, Symptom, Date*), which constitute the key information of the most critical defects.

System Acceleration

The **Time complexity** of model inference is: $O(Defect\ Count\ *\ Complaint\ Count\ *\ Vocabulary\ Size^2)$

In order to accelerate the inference, we reduce the vocabulary size by removing some less informative words. The Stanford NLP tagger is used to POS-tag a text, then only nouns, verbs, and adjectives are kept. We further conduct other pre-processing procedures (e.g., remove stop-words and punctuation, word transformation) to do further data filtering. Concurrent threads are also used for program acceleration. The processing speed increases around 10 times by these measures.

Evaluation

Both qualitative and quantitative evaluations are conducted for the evaluation of the proposed probabilistic model. In the qualitative step, the top defects of vehicle models are identified from the complaints on Honda vehicles. By reading the defect information, users can understand what kinds of issues occur according to these models. On the other hand, a quantitative test is done to evaluate the model's accuracy in terms of complaint clustering.

Dataset

In this experiment, we are using the complaint database of NHTSA, which has 1.13 million records in total, presenting complaints on various vehicle models. Each record has a number of attributes as introduced in Section 1.

For the qualitative test, 11560 complaints on 25 Honda's vehicle models produced between 2005 and 2009 were extracted from the database. In all the experiments, models with the same name but produced in different years are regarded as different models.

Regarding quantitative assessment, we synthesized 3 data sets by taking vehicle recall records as ground truth, to evaluate the performance of the proposed model on different vehicle brands. The process of synthesizing these data sets is:

- For a vehicle brand, find the top N critical recall records, from the vehicle recall database (DB), which have the most affected vehicles
- For each recall record, find corresponding complaints from the complaint DB by SQL
- Take the complaints of a recall record as a complaint partition

Table 1 shows the data sets produced for the quantitative evaluation.

Vehicle Brand	Number of Complaints	Number of Clusters
TOYOTA	1514	6
HONDA	1054	4
CHEVROLET	1675	5

Table 1. Data Sets for Evaluation

Qualitative Evaluation

ID	Model-Year	Defective Components	Symptom Words	Representative Complaints
1	ODYSSEY 2005	VISIBILITY	visor sun driver side break stay replace split windshield position	1. MELTING SUN VISOR. 2. PASSENGER SIDE SUNVISOR CAME APART AT THE SEAM. 3. DEFECECTIVE SUN VISOR OBSTRUCTING VISION.
2	PILOT 2005	POWER TRAIN	brake stop light pedal speed accelerate engine start gear problem	1. THIS VEHICLE SPUTTERS WHEN THE SPEED IS BETWEEN 35-40 MPH. 2. WON'T GO INTO SECOND GEAR. 3. TRANSMISSION SHIFTING PROBLEMS
3	ACCORD 2008	SERVICE BRAKES	tire brake rear mile replace wear problem pad noise air	1. 2008 HONDA ACCORD REAR BRAKES. 2. HONDA ACCORD 2008 BRAKES PROBLEM 3. REAR BRAKE WEAR

Table 2. Top Defects for 3 Honda Models

We used our algorithm to capture the top defects for Honda models produced between 2005 and 2009. The proposed method is able to identify the key information of each defect, including the most probable model-year, components, symptom words, and representative complaints. In addition, it can also find the most related defects of each vehicle model. Due to the limited space, Table 2 shows only the top defect of 3 Honda models. The defect information includes the flawed component, the symptom keywords, and some representative complaints. We can see the top defect of ODYSSEY 2005 is about the visibility component. Van owners are complaining about the defective sun visor. The transmission shifting problem is mostly reported by the PILOT 2005 drivers. In the case of ACCORD 2008, people criticize the service brake, which seems to be related to worn brakes.

From these defect samples, we can see the defect entities (component, symptom words, and representative complaints) reflect good consistency. Also, the defect information is able to indicate what kinds of problems are mostly complained about by people, and what types of symptoms automobiles have.

Quantitative Evaluation

In order to evaluate the accuracy of the proposed model, we take it as a clustering algorithm. For each complaint, we assign a cluster label to it using the highest probability of a relevant defect. In this way, we can build clusters of complaints according to their cluster labels. Then, the proposed model is evaluated by precision, recall, and F-Measure (Zaki and Jr 2014). Regarding baseline method, we use the K-Means clustering algorithm to identify different types of defects. The same vocabularies that have been used in PDM are applied as features in K-Means clustering for a fair comparison. The performance of the two clustering algorithms is shown in Table 3.

Data Sets	**PDM Performance**			**K-means Performance**		
Vehicle Brand	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Toyota	89.67%	87.10%	84.29%	78.90%	66.63%	68.19%
Honda	95.03%	91.83%	90.72%	90.05%	60.29%	61.91%
Chevrolet	87.80%	86.90%	84.53%	94.23%	58.95%	63.54%

Table 3. Clustering Performance of the Proposed Model and K-Means

From the results in Table 3, we can see:

- PDM outperforms K-Means in term of precision, recall, and F-Measure in most cases.
- The only exception is the precision of the Chevrolet data set. K-Means and lexical features seem to involve less noise in the Chevrolet case.
- The advantage of PDM is the most significant regarding the recall ratio.

Conclusion

Product issues are always concerning different groups such as customers, manufacturers, regulators, etc. People want to know not only what kinds of product issues occur most, but also detailed information about them. To address this problem, a probabilistic defect model is proposed in this paper, which tries to formulate the generation process of complaints. Complaints are taken as observations, while defect is the latent variable. By calculating the joint distribution of defect and inferring the parameters of entity distribution, we are able to identify the most critical defects and their most relevant entities, such as model, year, component, and symptom key words. An evaluation using the NHSTA vehicle complaint data proves the effectiveness of the proposed approach. The performance of the proposed model is better than the baseline method when evaluated by precision, recall, and F-Measure. Further, concurrent threads were applied to improve the efficiency of the proposed method.

In future work, we plan to improve the readability and the accuracy of the identified defects. The readability of the symptom words can be enhanced by replacing simple keywords with the main syntactic elements of the complaint sentences. In addition, the accuracy of defects might increase if we incorporate some prior distributions of the complaint entities into the graphical model.

REFERENCES

Abrahams, A., Jiao, J., and Fan, W. 2013. "What's Buzzing in the Blizzard of Buzz? Automotive Component Isolation in Social Media Postings," *Decision Support Systems* (55:4), pp. 871–882.

Abrahams, A., Jiao, J., Wang, G., and Fan, W. 2012. "Vehicle Defect Discovery from Social Media," *Decision Support Systems* (54:1), pp. 87–97.

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z., and Jiao, J. 2015. "An Integrated Text Analytic Framework for Product Defect Discovery," *Production and Operations Management* (24.6:6), pp. 975–990.

Aral, S., and Walker, D. 2011. "Creating Social Contagion through Viral Product Design: A Randomized Trial of Peer Influence in Networks," *Management Science* (57:9), pp. 1623–1639.

Bao, Y., and Datta, A. 2014. "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures," *Management Science* (60:6), pp. 1371–1391.

Blei, D., Ng, A., and Jordan, M. 2003. "Latent Dirichlet Allocation," *The Journal of Machine Learning Research* (3), pp. 993–1022.

Chen, Y., Ganesan, S., and Liu, Y. 2009. "Does a Firm's Product-Recall Strategy Affect Its Financial Value? An Examination of Strategic Alternatives during Product-Harm Crises," *Journal of Marketing* (73:6), pp. 214–226.

Chen, Y., and Xie, J. 2008. "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix," *Management Science* (54:3), pp. 477–491.

Dai, W., Brisimi, T., and Adams, W. 2015. "Prediction of Hospitalization due to Heart Diseases by Supervised Learning Methods," *International Journal of Medical Informatics* (84:3), pp. 189–197.

Ding, X., Liu, B., and Yu, P. 2008. "A Holistic Lexicon-Based Approach to Opinion Mining," *Proceedings of the 2008 International Conference on Web Search and Data Mining*.

Fan, W., and Gordon, M. 2014. "The Power of Social Media Analytics," *Communications of the ACM* (57:6), pp. 74–81.

Grimmer, J., and King, G. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization," *Proceedings of the National Academy of Sciences* (108:7), pp. 2643–2650.

Grimmer, J., and Stewart, B. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* (21:3), pp. 267–297.

Li, F., Han, C., Huang, M., Zhu, X., and Xia, Y. 2010. "Structure-Aware Review Mining and Summarization," *Proceedings of the 23rd International Conference on Computational Linguistics*.

Li, P., Jiang, J., and Wang, Y. 2010. "Generating Templates of Entity Summaries with an Entity-Aspect Model and Pattern Mining," *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Li, P., Wang, Y., Gao, W., and Jiang, J. 2011. "Generating Aspect-Oriented Multi-Document Summarization with Event-Aspect Model," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 1137–1146.

Li, Z., Wang, B., Li, M. and Ma, W.Y., 2005. "A probabilistic model for retrospective news event detection," *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval,* pp. 106-113.Liu, K., Xu, L., and Zhao, J. 2014. "Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-Ranking.," *ACL (1).*

Luo, X., Zhang, J., and Duan, W. 2013. "Social Media and Firm Equity Value," *Information Systems Research* (24:1), pp. 146–163.

Yan, Z., Xing, M., Zhang, D., and Ma, B. 2015. "EXPRS: An Extended Pagerank Method for Product Feature Extraction from Online Consumer Reviews," *Information & Management* (52:7), pp. 850–858.

Yang, B., and Cardie, C. 2013. "Joint Inference for Fine-Grained Opinion Extraction.," *ACL (1).*

Yu, Y., Duan, W., and Cao, Q. 2013. "The Impact of Social and Conventional Media on Firm Equity Value: A Sentiment Analysis Approach," *Decision Support Systems* (55:4), pp. 919–926.

Zaki, M., and Jr, W. M. 2014. *Data Mining and Analysis: Fundamental Concepts and Algorithms*.