Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Identifying Leading Indicators of Product Recalls from Online Reviews Using Positive Unlabeled Learning and Domain Adaptation

Shreesh Kumara Bhat, Aron Culotta
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
skumarab@hawk.iit.edu, aculotta@iit.edu

Abstract

Consumer protection agencies are charged with safeguarding the public from hazardous products, but the thousands of products under their jurisdiction make it challenging to identify and respond to consumer complaints quickly. In this paper, we propose a system to mine Amazon.com reviews to identify products that may pose safety or health hazards. Since labeled data for this task are scarce, our approach combines positive unlabeled learning with domain adaptation to train a classifier from consumer complaints submitted to an online government portal. We find that our approach results in an absolute F1 score improvement of 8% over the best competing baseline. Furthermore, when we apply the classifier to Amazon reviews of known recalled products, we identify safety hazard reports prior to the recall date for 45% of the products. This suggests that the system may be able to provide an early warning system to alert consumers to hazardous products before an official recall is announced.

1 Introduction[1]

The U.S. Consumer Product Safety Commission (CPSC) is charged with protecting the public from risks of injury from consumer products. The large volume of reports received by the CPSC make it difficult to investigate every complaint. For example, in FY2015 the CPSC issues 410 recalls; by comparison, the CPSC received 85,000 calls to their hot-line and 2,539 incidents submitted to their online portal.[2]

Given the large number of products under its jurisdiction, the CPSC faces a number of regulatory challenges, including (1) **discovery** to identify new potential hazards; (2) **triage** to prioritize potential hazards by severity; and (3) **notification** to inform consumers quickly of a potential hazard.

In this paper, we propose a system to help with these tasks by identifying product reviews on Amazon.com that indicate a potential safety or health hazard. The resulting system helps with discovery by identifying hazards that may not be submitted to the CPSC directly; it helps with triage by enabling complaints to be aggregated to identify high priority products; and it helps with notification by enabling consumers to be alerted immediately when hazardous reviews are posted on Amazon.

To train the text classification system, we use consumer complaints data uploaded to the CPSC portal. These "positive" instances are combined with thousands of unlabeled instances from Amazon.com reviews using Positive Unlabeled Learning (Li and Liu 2005). However, standard training algorithms underperform on this task, because these two data sources differ in systematic ways. To deal with this issue, we build on work in learning under dataset shift (Zadrozny 2004) to train a more accurate classifier. The resulting classifier identifies reviews mentioning safety hazards with an F1 score of 84%, an absolute improvement of 8% over the best baseline. Furthermore, we applied the classifier to reviews of known recalled products, and found that for 45% of the products, the system detected a review reporting a health or safety hazard prior to the recall date. This suggests that the system may be able to provide an early warning system to alert consumers to potentially hazardous products.

2 Data

Our goal is to build a text classifier to determine whether a product review on Amazon.com reports a potential safety or health hazard of a product. As such reviews are rare, it is difficult to construct a training set in the traditional way of annotating a random sample of reviews. Instead, we consider the consumer complaints database on the CPSC website *SaferProducts.gov*. We supplement this with a large set of unlabeled Amazon reviews to build the classifier using Positive Unlabeled learning. For validation, we consider two additional data sources: a small set of annotated Amazon reviews, and a set of products that were recalled by the CPSC over the past 10 years. We describe these data below.

CPSC Complaints Database The CPSC website *SaferProducts.gov* is a publicly searchable database of consumer submitted reports of hazardous products. For this paper, we focus on children products, since these tend to be the most vulnerable to health and safety hazards. We collected 2,010 complaints from the "Babies & Kids" category from *SaferProducts.gov*, from March 2011 – May 2016. The most popular categories are Cribs (407), Bassinets or Cradles (258)

[1] An expanded version of this paper is available at: http://arxiv.org/abs/1703.00518; replication files are at https://github.com/tapilab/icwsm-2017-recalls

[2] https://www.cpsc.gov/s3fs-public/FY15AnnualReport.pdf

and Diapers (209). When training the classifier, we assume that these 2,010 incident descriptions are positive examples (i.e., indicative of health or safety hazard). We refer to this as the **complaints database**.

Amazon Product Reviews We collect 915,446 Amazon reviews in the "Baby" category from the dataset introduced in McAuley et al. (2015), from August 2008 - July 2014.[3] We refer to this as the **reviews database**.

Labeled Review Data For validation, we manually annotated 448 Amazon reviews as to whether they report a hazardous or unsafe product. To construct this data, we combined uniform sampling with keyword search to identify possible positive examples (e.g., terms like "hurt" and "dangerous"). The final dataset contains 97 positive (hazardous) reviews and 351 negative (non-hazardous) reviews. A key challenge is distinguishing between reviews indicating a safety hazard and reviews that indicate more benign faults of the product. We refer to this as the **validation data**. An example labeled as hazardous is: "This item needs to be taken off the market. My son almost suffocated to death in this...". An example labeled as non-hazardous is: 'It's cheaply made. I washed it on the gentle cycle and it began to fall apart :("

Recall Database Finally, to explore the practical impact of this classifier, we collected a set of products that were recalled by the CPSC and had reviews in the reviews database. To do so, we first collected 6,741 recalled products from `cpsc.gov`. We used a semi-automated process to match each recalled product with an Amazon product in the reviews database. We identified 137 Amazon products that matched one of 47 recall records (some recalls affect multiple products). As this filtering makes clear, recalls are relatively rare events, so the data sparsity poses a challenge for typical machine learning training and validation workflows. This motivates our use of the complaints database to identify reviews indicating hazardous products.

3 Methods

Our goal is to train a classifier using the consumer complaints data to identify Amazon reviews indicative of a hazardous product. We have as input a set of positive examples from the complaints database and a set of unlabeled examples from the reviews database. Let $\mathbf{x}_i \in \mathbb{R}^k$ be the k-dimensional feature vector representing document i and $y_i \in \{0, 1\}$ be its class label, where 1 indicates a hazardous review. Then our input consists of positively labeled data $L = \{(\mathbf{x}_1, 1) \dots (\mathbf{x}_n, 1)\}$ of consumer complaints and unlabeled datas $U = \{x_1 \dots x_m\}$ of Amazon reviews.

This setting can be viewed as an instance of Positive Unlabeled learning (PU Learning (Li and Liu 2005)), since the training set consists of only positive and unlabeled instances. Below, we describe a simple baseline approach to this problem, identify a problem with this approach, then propose a new method that addresses this problem.

3.1 Baseline method

A simple approach to PU Learning is to assume that the unlabeled dataset U contains only negative examples; i.e., $U \triangleq \{(\mathbf{x}_1, 0) \dots (\mathbf{x}_m, 0)\}$. Of course, the unlabeled data may indeed contain positive examples; however, in our setting, hazardous reviews are rare in the reviews data, and so we expect the amount of label noise introduced to be low.

Furthermore, in this review domain, we also have the star rating of each review, which we can use to reduce the incidence of positive examples incorrectly annotated as negative examples in our training set. We expect reviews indicating safety hazards to have a low star rating. (While recalled products can have high *average* ratings, we expect individual reviews mentioning health hazards to have low star ratings.) So, we introduce a threshold τ when sampling negative examples from the unlabeled data; only instances with star rating greater than or equal to τ are selected. We also use a second parameter s indicating the number of negative examples to sample during training. We use logistic regression with L2 regularization as the baseline classifier. To handle class imbalance we weight each instance inversely proportional to its class frequency.

3.2 Proposed method: Informed prior

In addition to the small amount of label noise introduced by the baseline method, there is another, potentially more serious difficulty with the approach for this data. The problem stems from the selection bias in how the positive and negative examples are collected. Specifically, certain types of products like cribs, diapers, and night lights are over-represented in the complaints data relative to their prevalence in the reviews data. This leads to the inflation of coefficients related to these products — indeed, in the experiments below, we find that the terms "crib," "pampers," and "night light" are among the top ten coefficients for the positive class for the baseline classifier. This can lead to a number of false positives, in which reviews of these types of products are erroneously labeled as hazardous.

To address this problem, we build on work in learning under dataset shift (Zadrozny 2004). Our approach modifies the feature representation so that terms that are strongly predictive of the positive class in the unlabeled dataset have larger feature values than terms that are less predictive. Of course, we do not know the true labels in the unlabeled data; we instead use the baseline classifier to estimate them.

Our approach begins by fitting the baseline classifier, as defined in the previous section. We refer to the training data for this classifier as the **baseline training set**. We then apply this classifier to predict the labels for all unlabeled reviews in the Amazon review data; we refer to this as the **predicted reviews data**. Based on the examples above ("crib", "pampers," etc.), the key observation of our approach is that certain word features may be strongly associated with the positive class in the original training data, but may be weakly associated with the positive class in this predicted reviews data. For example, in one experiment below with $\tau = 5$ and $s = 20,000$, we find that in the baseline training set, 91% of documents with the term "pampers" were annotated as positive examples (i.e., were from the complaints data). However, in the predicted reviews data, only 2% of documents containing the term "pampers" were predicted to be positive examples by the baseline classifier. So, our motivation is to

[3] `http://jmcauley.ucsd.edu/data/amazon/`

increase the importance of features that are strongly associated with the positive class in the predicted reviews data. We do this by modifying the value for features proportional to their class conditional probability in the predicted reviews data, as described next.

Let $\hat{U} = \{(\mathbf{x}_1, \hat{y}_1) \dots (\mathbf{x}_m, \hat{y}_m)\}$ be the predicted reviews data; i.e., all Amazon reviews and the corresponding class labels predicted by the baseline classifier. Let $\theta_j \in \mathbb{R}$ be the coefficient in the baseline model associated with word feature j, and let $x_i^j \in \{0, 1\}$ be the binary feature value for feature j in document i. For each term feature, we compute the smoothed class conditional probability according to the predictions in $\hat{U}$. Let n_{jc} be the number of documents containing feature j that have been assigned label c by the baseline classifier. Then we can define the conditional probability with Laplacian smoothing as:

$$p(y = 1|x^j = 1) = \frac{1 + n_{j1}}{2 + n_{j1} + n_{j0}} \triangleq p_{j1}$$

and similarly for p_{j0} for class 0.

Let F^+ be the set of features with positive coefficients in the baseline classifier, and similarly for F^- for negative coefficients. We will use p_{j1} to transform the feature values for F^+, and p_{j0} to transform the feature values for F^-. In order to have the transformation be in the same scale for each class, we first normalize the conditional probabilities: $\hat{p}_{j1} = \frac{p_{j1}}{\sum_{j' \in F^+} p_{j'1}}$, and similarly for $\hat{p}_{j0}$. To construct suitable feature values, we want to shift these values to have a mean of 1 and be non-negative, which we can do by multiplying each value by a constant factor ρ, the ratio of the number of features to the sum of the values $\hat{p}_{jc}$:

$$\rho = \frac{|F^+| + |F^-|}{\sum_{j \in F^+} \hat{p_{j1}} + \sum_{j \in F^-} \hat{p_{j0}}}$$

Finally, for all instances in the training and unlabeled data, we replace the value of feature j with the factor $(\rho * \hat{p}_{j1})$ if $j \in F^+$ or with $(\rho * \hat{p}_{j0})$ if $j \in F^-$.

As an example from one of the experiments below, the feature value for the bigram "very dangerous" is increased to 17.4, because 29% of documents in the unlabeled data containing "very dangerous" were classified as positive by the baseline classifier, the second highest rate of all features. Conversely, the term "crib" only has a feature value of 2.1, because only 3% of documents in the unlabeled data containing "crib" were classified as positive. This is particularly notable given that the baseline model assigns a higher coefficient to "crib" (1.34) than to "very dangerous" (0.55).

4 Experimental Results

We tokenize 915,446 Amazon reviews, retaining unigrams and bigrams that appear in at least 50 reviews and no more than 95% of all reviews, resulting in 136,160 total features. We represent each review as a binary feature vector. Using this pruned feature set, we then vectorize the 2,010 messages in the complaints dataset, as well as 448 labeled Amazon reviews in the validation data. For evaluation we use Precision/Recall/F1 as well as area under the ROC curve.

Model	Rev. Thresh. (τ)	AUC	F1	Prec	Rec
inf. prior	5	**97.0**	**84.3**	85.8	**82.8**
inf. prior	4	96.4	82.7	86.8	79.0
inf. prior	3	96.3	82.1	**87.6**	77.3
baseline	5	96.1	75.3	72.8	78.0
baseline	4	95.9	74.8	73.7	75.9
baseline	3	95.7	76.4	78.4	74.6
baseline	none	94.0	70.0	79.0	62.9

Table 1: Comparison of the baseline classifier with our informed prior method on the validation data.

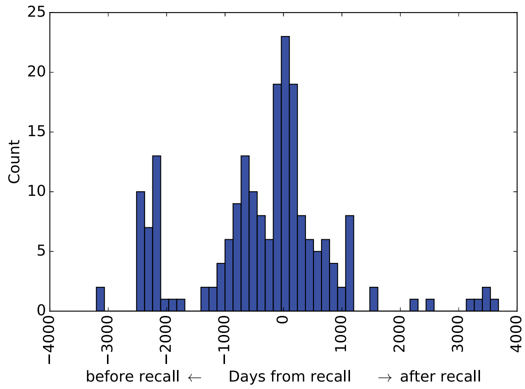

Figure 1: Histogram of when each of the identified hazardous reviews was submitted to Amazon relative to the date that the product was recalled.

Since the baseline training set is constructed by sampling s random reviews with rating $\geq \tau$ from the unlabeled Amazon review data, we average the results of three trials. We consider three empirical questions in the remainder of this section.

(1) How does our informed prior approach compare to the baseline classifier? The primary classification results on the validation data are shown in Table 1. To start, we fix $s = 20,000$. We observe that across all performance measures the informed prior method produces more accurate results than the baseline. To better understand these results, we examined the terms with the highest positive coefficients for the informed prior and baseline classifiers, using $s = 20,000$ and $\tau = 5$. We found that the baseline model has many words that are likely due to sampling bias, such as "pampers", "crib," "night light," "gate," and "model." On the other hand, the informed prior model gives higher weight to features such as "very dangerous," "emergency room," and "is unsafe." Recall that both models are fit using the same training instances; the only difference is that feature values are increased for terms estimated to be predictive of the positive class in the unlabeled data. We also note that terms like "cpsc" and "recalled" arise from that fact that some reviews either discuss a pending or past recall of a product, or indicate that they have concurrently posted a complaint to the CPSC database.

(2) How do the parameters τ and s affect accuracy? Table 1 lists results as we change τ, the review threshold used when sampling examples from the unlabeled data to serve as negative training instances. We can see that increasing this threshold can greatly improve the recall of the classifier, while sometimes reducing precision. However, the overall AUC increases as τ increases. We conjecture that the boost in recall is in part because by removing reviews with low ratings, we remove from the training set reviews with negative sentiment that are labeled as non-hazardous reviews.

To investigate the impact of s, we measured the F1 score of the informed prior classifier ($\tau = 5$) as s increases. We found that generally accuracy is stable for s in the range $5,000 - 40,000$. For values outside that range, class imbalance begins to negatively impact accuracy.

(3) How often does the classifier identify potential product hazards before a recall is issued for a product? Using the best classifier from Table 1 (informed prior; $\tau = 5; s = 20,000$), we next predict the label for the reviews of the Amazon products identified as being part of a CPSC recall. After filtering products with fewer than 10 reviews, we are left with 7,318 reviews from 86 products, of which 204 reviews were predicted to report a safety hazard.

To investigate the ability to provide consumers with a quicker notification of potential hazards, Figure 1 shows a histogram of when each of the identified hazardous reviews were submitted to Amazon relative to the product recall date. Overall, for 45% of the recalled products, the classifier identified at least one hazardous review prior to the recall date. A manual analysis reveals that several complaints described in the Amazon reviews are also mentioned in the reason for the recall posted by the CPSC (e.g., a review complaining about the front wheel assembly of a stroller was followed by a recall based on the same issue). There are some outliers appearing years before the recall date; this can happen when a recall is issued because of stores that continue to sell merchandise that had already been recalled. Additionally, there are reviews found well after the recall date, which can occur for products that have been discontinued on Amazon, but still have a page on which users can submit reviews.

Taken together, these results suggest that there is an opportunity to mine Amazon reviews to provide earlier warnings to consumers about potentially hazardous products, as well as to prioritize complaints posted on Amazon for potential examination for safety hazards.

5 Related Work

To the best of our knowledge, this is the first published system to identify product health and safety hazards from online reviews with no manual human annotation required. Additionally, our time-series experiments indicate that these reviews can be identified prior to the product recall date.

Very recently, Winkler et al. (2016) used a keyword based approach to identify online reviews that report injuries from toy products. In addition to the manual effort required to curate the keyword list, the approach appears to produce low precision rates (9-44%, depending on subcategory). Of the top 100 identified reviews, only sixteen mentioned an injury. The authors apply the same approach to detect defects in dishwashers, with similar precision values (Law, Gruss, and Abrahams 2017). In contrast, our proposed approach fits a statistical classifier with no human intervention required, resulting in $> 85\%$ precision and $> 80\%$ recall.

Other recent work has identified vehicle defects in consumer reviews using standard text classification, with accuracies ranging from 62%-77% (Abrahams et al. 2015). However, in many domains it is not feasible to annotate sufficient messages to use standard supervised learning. Additionally, Zhang et al. (2015) built an unsupervised approach to clustering vehicle defects by subcategory. Such a method may serve to complement our present work by providing more fine-grained clusters of reviews by hazard type.

6 Conclusion

We have presented a classification system to identify product reviews on Amazon.com that indicate a health or safety hazard. The classifier is trained without any additional human annotation or intervention by using the consumer complaint records submitted to SaferProducts.gov. To deal with data selection bias, we introduced a new domain adaptation approach that is easy to implement and results in an 8% absolute increase over the best competing baseline. An analysis of the historical reviews of recalled products indicates that the system can identify potential safety hazards well before the recall is issued.

Acknowledgments

This research was funded in part by the National Science Foundation under awards #IIS-1526674 and #IIS-1618244.

References

Abrahams, A. S.; Fan, W.; Wang, G. A.; Zhang, Z. J.; and Jiao, J. 2015. An integrated text analytic framework for product defect discovery. *Production and Operations Management* 24(6):975–990.

Law, D.; Gruss, R.; and Abrahams, A. S. 2017. Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications* 67:84–94.

Li, X.-L., and Liu, B. 2005. Learning from positive and unlabeled examples with different data distributions. In *European Conference on Machine Learning*, 218–229.

McAuley, J.; Targett, C.; Shi, Q.; and van den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 43–52. ACM.

Winkler, M.; Abrahams, A. S.; Gruss, R.; and Ehsani, J. P. 2016. Toy safety surveillance from online reviews. *Decision support systems* 90:23–32.

Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, 114.

Zhang, X.; Niu, S.; Zhang, D.; Wang, G. A.; and Fan, W. 2015. Predicting vehicle recalls with user-generated contents: A text mining approach. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, 41–50. Springer.