"What's wrong with this product?" - Detection of product safety issues based on information consumers share online

Max Fuchs*
Maastricht University
Maastricht, Netherlands

Amit Jadhav*
Maastricht University
Maastricht, Netherlands

Advaith Jaishankar*
Maastricht University
Maastricht, Netherlands

Caroline Cauffman
Maastricht University
Maastricht, Netherlands

Gerasimos Spanakis
jerry.spanakis@maastrichtuniversity.nl
Maastricht University
Maastricht, Netherlands

ABSTRACT

With the widespread use of e-commerce, proper oversight and regulatory compliance become increasingly difficult, if not impossible, resulting in a heightened risk of harm to consumers from unsafe products. In this paper, we explore how online consumer reviews can be utilized to identify hazardous products that have previously been flagged in the European Union Safety Gate reports. Our research presents a general framework that can be beneficial for regulatory authorities, as well as a specific application to consumer electronics. We contribute a dataset of 3000 reviews of electronic products, 755 of which reference hazardous products, and conduct classification baselines, achieving an AUC of up to 80% with room for improvement. Furthermore, we discuss the legal basis for annotation and potential issues that may arise. Our proposed methodology and dataset are valuable resources for regulatory authorities in the European Union and provide evidence of the effectiveness of digital surveillance in protecting consumers.

CCS CONCEPTS

• **Applied computing** → **Investigation techniques**; • **Information systems** → **Data mining**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

datasets, consumer protection, online reviews, classification

ACM Reference Format:
Max Fuchs, Amit Jadhav, Advaith Jaishankar, Caroline Cauffman, and Gerasimos Spanakis. 2023. "What's wrong with this product?" - Detection of product safety issues based on information consumers share online. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3594536.3595145

*Authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICAIL 2023, June 19–23, 2023, Braga, Portugal
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0197-9/23/06…$15.00
https://doi.org/10.1145/3594536.3595145

1 INTRODUCTION

The increasing use of Internet has made it possible to purchase products online which on the one hand facilitates consumers (ease of shopping, price comparisons etc.) but on the other hand has several drawbacks (increasing cases of fraud etc.) [6]. More specifically, the wide availability of online shops (across different countries) makes it hard for the relevant regulatory authorities to monitor market conditions. Under European Union (EU) laws [15], products that are for sale in EU countries are subject to specific safety regulations and are often tested before being allowed to be brought onto the market. The lack of proper oversight due to the massive size of the Internet and the dispersion of online shopping has led to potentially unsafe products being sold and potentially harming consumers [15]. Such harms include choking hazard, exposure to poisonous materials, burns and even death [1].

The European Union Safety Gate (EUSG from now on) portal[2] provides an overview of products types have been reported as hazardous from various European regulatory authorities. This short paper focuses specifically on electrical appliances and equipment, which are also commonly purchased online compared to e.g. cars. An example of a Safety Gate report can be seen in Figure 1.

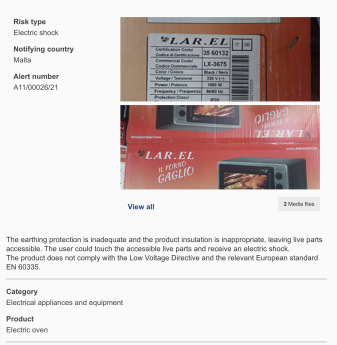

Figure 1: Example of a report from the European Union Safety Gate (EUSG) portal

Growth of e-commerce has led to the growth of the amount of online information on purchases in the form of either structured

[1] https://www.wired-gov.net/wg/news.nsf/articles/Safety+Gate+Motor+vehicles+and+toys+top+the+list+of+dangerous+nonfood+products+this+year+26042022143300?open
[2] https://ec.europa.eu/safety-gate-alerts/

reviews on a variety of websites or unstructured social media posts. The vast availability of such types of data has enabled researchers to conduct different studies involving product reviews by extracting information from the data available online. Such reviews can potentially include severe safety hazards (overlapping with some of the cases reported in Safety Gate), like for example can be seen in Figure 2.


Tubborg

★☆☆☆☆ **Safety risk - short circuit**

Reviewed in Italy on 13 February 2019

Style Name: Contact grill 3-in-1. | Verified Purchase

The product has literally exploded during use after a few months from purchase for a short circuit on the power cord (from the photos it is evident the cleft on the geneatasi cable following the short internal circuit).
Big scare (he made a nice "bang") and real risk to security, luckily the lifesaving of the apartment intervened.

Figure 2: Example of a review from the Amazon website

Therefore, the following question is being raised: Is it possible to detect products purchased online (in this case electronic ones) that might lead to safety hazards based on online data (more specifically product reviews)? This question is relevant for all different stakeholders: regulatory authorities (such as consumer protection agencies) since it gives them the possibility to easily filter online reviews and audit potentially hazardous products and consumers since they can be better protected against unsafe products.

Our approach considers first reports from EUSG portal in order to construct a list of terms ("smoke words") that have been used to describe hazardous products. This list is used as a first filtering of the online reviews in order to detect potentially unsafe products. Subsequently, a set of 3000 reviews is annotated for potentially unsafe products. Then, we compare two methodologies (one that uses the Correlation Coefficient (CC) Score and one that uses Machine Learning (ML) classifiers) that on the basis of the annotated data can detect whether a review refers to a potentially unsafe product or not.

The main contributions of this paper are twofold: First, we provide an annotated dataset of 3000 reviews of electronic products on the basis of whether there is a reference to potential hazards along with benchmark results on the subsequent classification task. The results of this paper underline the importance of properly annotated data and the search for it online has not been of great success. Secondly, we provide a general framework (Figure 3) that regulatory authorities in the EU (and beyond) can use so that they can easily navigate online reviews (either on e-commerce websites or social media) that might contain potentially hazardous products.

2 RELATED WORK

There have been different approaches in literature to address the problem of detecting reviews of hazardous products using different sources of data (social media, online customer reviews, discussion threads in consumer forums etc.). In [1] and [17] authors curate domain-specific (for cars and toys respectively) "smoke words" (words that are more prevalent than others in safety/defect issues) that can be used to identify defect products in online reviews (either forums or reviews). Similarly, in [13] and [2] authors follow a similar approach but try to automate the smoke word list creation through the reviews and their application domains are baby cribs and joint and muscle pain relief treatments respectively.

Lately, researchers have been experimenting with different domains but also different machine learning techniques (e.g. [12] compare Logistic Regression, Decision Trees and Neural Networks) while [9] use Recurrent Neural Networks (RNN) for sentiment analysis of reviews so as to identify negative ones and Latent Dirichlet Allocation (LDA) as a way to retrieve a summary of key defect insight words from these reviews). LDA has also been used in [18] in order to identify domain-specific knowledge about product issues but due to the unsupervised approach, their results are difficult to assess.

Similar to our approach, [3] compiled product reviews from Amazon.com along with consumer complaints from SaferProducts.gov complaints and product recall descriptions. Their domain of application is baby toys and they provide a comprehensive labeling of the safety issues in online reviews, however their performance is quite low (precision of 60% in detecting unsafe products).

Our work draws a general framework that can be used to detect products with safety issues based on consumer reviews that are shared online. The overview of this framework can be seen in Figure 3. In the following chapters, we will describe the specific components of this framework.

3 DATA

In this chapter we describe the data sources (namely the EUSG reports and the Amazon reviews) as well as the subsequent analysis conducted in each. We use the Safety Gate reports to extract a smoke term list for describing hazardous products and then we use this list as a basis for annotating Amazon reviews on the basis on whether the review contains a hazard or not.

3.1 Europen Union Safety Gate reports

The European Union Safety Gate (EUSG) is used by European Union market surveillance authorities to register unsafe products, including those that present a risk to the health and safety of consumers. The platform is used by the competent authorities across the EU as a single alert system for dangerous consumer products (excluding food, pharmaceutical and medical products). An example of such an entry can be found in Figure 1.

As already mentioned this paper focuses on electronics, a category which provided us around 3000 reports about unsafe products. These reports were analyzed to determine what are the most common terms found in these reports. The most popular sub-categories of hazards can be found in Table 1.

We define a "smoke term" as a word or phrase highly correlated to textual content of interest [10], thereby, its presence is most of the times indicative of the content of interest. For the particular case study, the smoke word list was created using the categories of hazards and the reports on hazards from the ECSG. More specifically, the list of all categories (a part of which can be see in Table 1) as a whole was included in the initial smoke word list. In order to extract the smoke words from the reports, IDF word weighting [16] was used across all reports and the top-words were inspected. We further process the list by removing stop-words (using the common English list of NLTK) and by including synonyms (using WordNet). After this curation, the final list contains 41 terms and can be found in Table 2.

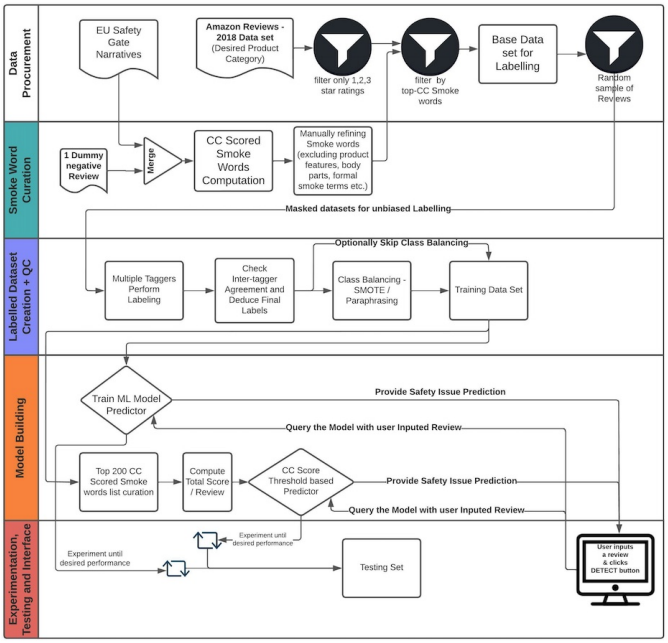

Figure 3: Framework for Safety Issue detection across multiple product categories transacted on e-commerce websites

Category	Frequency
Electric shock	1789
Electric shock, Fire	350
Fire	220
Burns	96
Burns, Electric shock, Fire	94
Environment	75
Burns, Fire	75
Burns, Electric shock	67
Injuries	59
Electric shock, Injuries	18

Table 1: Top 10 types of hazard types for electronic products

1	dangerous	14	damage	27	creepage	40	smoke
2	asphyxiation	15	plug	28	temperature	41	hazard
3	burns	16	electrocuted	29	causing		
4	chemical	17	comply	30	contacts		
5	choking	18	shock	31	unsafe		
6	cuts	19	voltage	32	safety		
7	electric	20	live	33	thermal		
8	fire	21	insulation	34	wire		
9	health	22	power	35	reported		
10	injuries	23	cause	36	plugs		
11	risk	24	serious	37	incompliant		
12	strangulation	25	overheat	38	short		
13	suffocation	26	insufficient	39	wiring		

Table 2: Smoke word list generated with TF-IDF and manual revision

3.2 Amazon reviews

As source for the reviews on consumer electronics, the choice was made for a historical dataset that contains 20,994,353 historical reviews on consumer electronics [14]. We decided to keep reviews with 1-, 2- or 3-star ratings (in total 5,082,583 reviews) since they are most likely to contain unsafe product issues. Applying smoke term list of Table 2 on the filtered dataset of approx. 5 million reviews ended up producing 875k reviews that contain one or more smoke words

3.2.1 Data Annotation. The filtered dataset contains potentially hazardous reviews, however in order to establish that, we decided to manually annotate the data using the following process. Two copies of six random samples of 500 reviews each, were assigned to 6 human annotators for labeling such that each sample of 500 reviews is independently annotated by 2 individuals. Annotators could label the review as "hazardous" or "non hazardous" or "unsure" and additionally they could indicate which words believed they contributed to the context of the review being hazardous. Annotators were law students that could assess which product reviews demonstrated a potential hazard from a legal point of view and which were merely about discomfort to the person or defect in the product being used, etc. Even though the discomfort can be annoying, sometimes it is not immediately a threat to the safety of the consumer, thus there is no legal ground for the product to be declared unsafe. This is why this nuance in the labeling is necessary.

A total of 3000 reviews were evaluated, and to determine the level of agreement regarding the labeling of each review, Cohen's Kappa scores [4] were calculated for the label chosen by the two individuals. The results showed a Cohen's Kappa score of 0.50 (n = 3000), indicating "fair" to "good" agreement. Out of the 3000

reviews, 36 were marked with a blank label because at least one annotator was uncertain, while 2829 reviews were agreed upon and 135 reviews had disagreements. Overall, the dataset had a 95% agreement rate.

A third annotator was used for disagreements (135 cases) and for reviews where one or both annotators had assigned a blank label (36 cases). This annotator also checked a random sample of 50 reviews and found that all the labels were correctly assigned. After performing this third level of labeling, the label assigned by the majority (i.e. two out of the three annotators) was considered as the final label. We further discuss annotation issues in Section 5.

4 EXPERIMENTS AND RESULTS

In this section we will present the models used for detecting safety issues with products based on consumer reviews and their comparative results.

The final dataset contains 2245 negatively labeled reviews against 755 positively labeled reviews (i.e. non-hazardous vs. hazardous). Results presented here are on the initial unbalanced dataset, since techniques for balancing the dataset were inconclusive. The dataset is split with stratified sampling to 1794 data points (1340 negative, 454 positive) used for training and 1206 data points (905 negative, 301 positive) used for testing.

4.1 Approach 1: Correlation Coeffecient (CC)-Score

Based on literature ([17]), we first experimented with the Correlation Coefficient (CC) which has been shown to be an effective metric [8] in the problem examined. The advantage of this method is that is requires no model to build and little human intervention rather than a list of possible smoke terms that highly contribute to the description of a safety issue. The CC-score is a variation of the Chi-Square measure and has been proposed so as to deal with issues of including non-relevant words when discriminating two classes (like in our case). The formula for computing the correlation coefficient of any given word is given below:

$$CC = \frac{\sqrt{N} \times (AD - CB)}{\sqrt{(A+B) \times (C+D)}}$$

where A is the number of relevant documents containing the word, B is total number of non-relevant documents containing it, C is the total number of relevant documents not containing the word and D is the total number of non-relevant documents not containing the word. A, B, C and D together sum to N.

The CC-score was computed on the reviews of the training set and we leave the testing set for checking the success of the approach.

We pick the 200 words with the highest CC-score on the training set. The choice of 200 is based on the manual observation that the relatedness of words to hazardous issues started to diminish. In order to assess the method, each review was assigned a score based on the accumulated CC-score of all instances of this 200 words list. Contrary to previous literature (e.g. [13]), we normalize the score by the length of the review, since otherwise the final score will be highly affected by how many words each review has.

Furthermore, we did experiment with the threshold of the normalized CC-score, above which a review will be classified as hazardous. The results of this experiment are presented in Table 3. Method achieves best AUC and accuracy for a threshold of 10, however given that recall (which might be prioritized for this problem so as not to exclude any hazardous reviews) seems to be slightly better for lower thresholds, one might choose a CC-score threshold of 8, while not sacrificing much precision.

Threshold	Precision	Recall	F1-score	Accuracy	AUC
6.5				0.75	0.75
Class 0	0.83	0.61	0.71		
Class 1	0.69	0.88	0.78		
8:				0.76	0.76
Class 0	0.81	0.67	0.73		
Class 1	0.72	0.84	0.78		
10:				**0.77**	**0.77**
Class 0	0.78	0.74	0.76		
Class 1	0.75	0.79	0.77		
12:				0.76	0.76
Class 0	0.76	0.80	0.77		
Class 1	0.78	0.73	0.75		
14:				0.74	0.74
Class 0	0.70	0.83	0.76		
Class 1	0.80	0.65	0.72		

Table 3: CC-score experiment results

4.2 Approach 2: Training classifiers

We focused on both sparse and dense representation models and different classifiers. Details of the classifiers are presented below and results on the test set are presented in Table 4.

Method	Precision	Recall	F1-score	Accuracy	AUC
BiLSTM with custom embeddings:				0.54	0.46
Class 0	0.94	0.56	0.70		
Class 1	0.05	0.36	0.08		
BiLSTM with Glove-100 embeddings:				0.57	0.50
Class 0	0.94	0.58	0.72		
Class 1	0.06	0.41	0.01		
BiLSTM with GoogleNews embeddings:				0.69	0.49
Class 0	0.94	0.72	0.81		
Class 1	0.05	0.27	0.09		
SVM with custom embeddings:				0.93	0.77
Class 0	0.98	0.95	0.96		
Class 1	0.41	0.61	0.49		
Logistic Regression with custom embeddings:				**0.91**	**0.79**
Class 0	0.98	0.92	0.95		
Class 1	0.35	0.68	0.46		

Table 4: Classifier results

Custom embeddings were trained on our dataset using the skip-gram architecture with a context window size of 5 and embedding size of 300. As an alternative, we used pre-trained embeddings, namely the 100-dimensional GloVe ones (pre-trained on Wikipedia) and the 300-dimensional Google News ones.

Out of all the models (as seen in Table 4) we can see that SVM and Logistic Regression models trained on custom embeddings have the highest precision, recall and AUC.

5 DISCUSSION

The false negative cases for the 65% prediction probability threshold (based on the LR model on custom embeddings) and for the normalized CC-score method (with threhsold 8) were manually

inspected. There is a great overlap between the mistakes made by the two methods and they both seem to be driven by the same reasons. First of all, some errors were caused by annotation issues. Despite that fact that we picked law students for annotating (on legal grounds), it is not easy to determine with absolute certainty whether a product will legally qualify as safe or unsafe based on the information available in the review. In the EU, product safety is guaranteed by a mix of public law regulation and private law liability rules. The public law approach is to determine criteria products must meet in order to be made available on the market. These criteria may be determined by international, EU or national law. At the EU level and in the absence of more specific legislation, these criteria are determined by the General Product Safety Directive [5] and include voluntary standards, and reasonable consumer expectations regarding safety.

The private law approach is to hold producers liable for harm caused by defects in their products. In most EU Member States, the general rules on non-contractual liability will apply. In addition, the EU Product liability directive [7] requires Member states to introduce a no-fault product liability. In particular, producers of defective products are to be held liable for harm caused to persons, and - subject to a minimum threshold of 500 euro - for damage to items of property (other than the defective product) of a type that is 'ordinarily intended for private use or consumption, and used by the injured person mainly for such purpose'. For the purpose of this rule, a product is considered defective when it 'does not provide the safety which a person is entitled to expect, taking all circumstances into account' [11].

Based on the consumer reviews, annotators could not assess whether the products in question satisfied all criteria required by public law rules. Neither were they aware of any potential negligence which might affect the application of the rules on non-contractual liability, or of any specific circumstances surrounding the offer, which might affect the safety assessment under the harmonized product liability rules. The annotators therefore limited themselves to assessing to the best of their ability, whether the goods provided the safety a person is entitled to expect.

Furthermore, some errors occur due to specific language synonymy issues (e.g. Amazon Fire product and fire as a hazard word). However, in most false negative cases, the reviews contained words which were strong indicators of hazards but comparing these with the smoke word list, it was found that these words are present but only in a different form (e.g. synonym). That gives rise to an interesting future direction, which would be to "match" the vocabulary used by online reviews and the one used in formal channels and how to better create smoke-word lists.

Regardless of these limitations, this paper introduces a new framework (Figure 3) for identifying hazardous products from online reviews and initial experiments reached an AUC score of approx. 80%, therefore we believe that our research will motivate researchers to create high quality labeled datasets, annotated on the basis of legal rules and norms is necessary. We hope that this discussion will spark new ideas for collaborations and potentially solve consumer protection issues, both for consumers but also for the regulatory authorities that need to oversee digital commercial markets.

ACKNOWLEDGMENTS

We would like to express our gratitude to the students from the Faculty of Law of Maastricht University for their annotation efforts.

REFERENCES

[1] Alan S Abrahams, Jian Jiao, G Alan Wang, and Weiguo Fan. 2012. Vehicle defect discovery from social media. *Decision Support Systems*, 54, 1, 87–97.

[2] David Z. Adams, Richard Gruss, and Alan S. Abrahams. 2017. Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International Journal of Medical Informatics*, 100, 108–120. 108.

[3] Graham Bleaney, Matthew Kuzyk, Julian Man, Hossein Mayanloo, and Hamid R Tizhoosh. 2018. Auto-detection of safety issues in baby products. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 505–516.

[4] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 1.

[5] European Commission. 2002. Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on general product safety. L 11/4. *Official Journal of the European Communities*.

[6] Feng Ding, Jiazhen Huo, and Juliana Kucht Campos. 2017. The development of Cross border E-commerce. In *International Conference on Transformations and Innovations in Management (ICTIM 2017)*. Atlantis Press, 487–500.

[7] Council Directive. 1985. Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products. *Official Journal L*, 210, 07/08, 0029–0033.

[8] Weiguo Fan, Michael D Gordon, and Praveen Pathak. 2005. Effective profiling of consumer information retrieval needs: A unified framework and empirical comparison. *Decision Support Systems*, 40, 2, 213–233.

[9] Titus Hei Yeung Fong, Shahryar Sarkani, and John Fossaceca. 2021. Auto Defect Detection Using Customer Reviews for Product Recall Insurance Analysis. *Frontiers in Applied Mathematics and Statistics*, 38.

[10] David M. Goldberg, Richard J. Gruss, and Alan S. Abrahams. 2022. Fumeus: A family of Python tools for text mining with smoke terms. *Software Impacts*, 12, 100270.

[11] Geraint Howells, Christian Twigg-Flesner, and Thomas Wilhelmsson. 2017. *Rethinking EU consumer law*. Taylor & Francis.

[12] Darren Law, Richard Gruss, and Alan S. Abrahams. 2017. Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications*, 67, 84–94.

[13] Vaibhav Mummalaneni, Richard Gruss, David M Goldberg, Johnathon P Ehsani, and Alan S Abrahams. 2018. Social media analytics for quality surveillance and safety hazard detection in baby cribs. *Safety science*, 104, 260–268.

[14] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 188–197.

[15] United Nations Conference on Trade and Development. 2018. *Trade and Development Report 2018*. United Nations.

[16] Gerard Salton. 1983. Introduction to modern information retrieval. *McGraw-Hill*.

[17] Matt Winkler, Alan S Abrahams, Richard Gruss, and Johnathan P Ehsani. 2016. Toy safety surveillance from online reviews. *Decision support systems*, 90, 23–32.

[18] Xuan Zhang, Zhilei Qiao, Aman Ahuja, Weiguo Fan, Edward A Fox, and Chandan K Reddy. 2019. Discovering product defects and solutions from online user generated contents. In *The World Wide Web Conference*, 3441–3447.