Master Research Project: "What's wrong with this product?" Detection of product safety issues based on information consumers share online

Kilian Kramer

June 2022

1 Abstract

The ability to purchase products online has provided many advantages for consumers in the recent years. With these advantages however, there are also drawbacks. One of these drawbacks, is the fact that ordering online and especially from foreign markets, makes it hard for oversight to properly be carried out by governmental bodies. Products, produced in the EU are subject to safety regulations and are often tested before being allowed to be brought onto the market. It is known that, because of the lack of oversight, products are purchased and used that can cause harm to humans. The amount of harm ranges from choking hazard, poisonous materials, burns and even death [1]. Goal of this paper is to give an insight on how to utilize machine learning to tackle this problem and make it easier to detect such hazardous products.

2 Problem Statement and Research Questions

Previous work that aims at solving a similar problem has been conducted. This work was however conducted on a specific categories of consumer products. One such work is in pharmaceutical products. The methods used for this are not directly applicable to the now given problem, but are referenced and used as inspiration.

The previous work uses social media in order to detect side effects of pharmaceutical products. The interesting thing about this, is that the side effects became apparent through social media earlier than through official channels.

Given this approach, it might be interesting to investigate a similar strategy. The large amount of online available data by many retailer websites can be used for this. This work will focus on using online reviews and apply modern deep learning techniques to classify these reviews. To delimit the scope of the project, it will focus only on electronic articles, but within electronics articles on several subcategories. For future work the pipeline of the project will be documented in a manner thus it can be applied on other product categories equally.

The paper will focus on the following questions:

1. What are the necessary steps to obtain and prepare online data such in can be utilized to perform advanced machine learning techniques?
2. How good will perform a modern text classifier like a bidirectional LSTM to detect hazards in reviews?
3. Which difficulties are we facing by using this approach and this domain?

3 Related Work

Literature points to multiple approaches taken to address the problem at hand, to identify the product safety issues. Researchers have worked on diverse sources of data such as social media data, online customer reviews' data and discussion threads in consumer forums among others. The array of natural language processing techniques used in various studies is also equally diverse. This section is an attempt to highlight and build upon the previous work.

[2] suggest, product safety issues are associated with highly negative sentiments of customers reactions on social media.[3] in an approach contrasting to [2], emphasize on the importance of smoke words over highly negative sentiments to track product defects due to the loss of objective diversity in the customer reviews when depending solely on sentiments. Rightfully so, customers who post a review about a defect can be equally keen to give out an objective description of the defect, than just uttering negative words to express their frustration.[4] have combined smoke and sparkle terms with sentiment analysis by first tagging the data manually and using the tags thus created for sentiment analysis to detect the defects in dishwasher appliances. One limiting factor to relying heavily on such domain specific lexicons as smoke words is that the tying together of these lists involves a dependency on subject matter experts and also an equally considerable amount of time. Such smoke terms, albeit efficient, but attributing to their domain specific nature, can also have limitations in scalability across domains. A list of these terms, for instance, which facilitates an accurate issue detection in a specific domain of products can perform badly across another domain of products. More recent literature utilizes topic modelling using Latent Derichlet Allocation (LDA) [5], a combination of artificial neural networks with LDA [6] and probabilistic graphical models (PGM) [7] to extract detailed

information about the products while retaining highly relevant product feature information to trace the defects which is not as effectively possible using Sentiment analysis and Smoke words as standalone approaches.

[8] explains the need to look at emoticons or Emojis contained as a part of social media textual data as these can also be strong indicators of the instantaneous emotions of the person authoring the review. It is also a possibility that the rating given alongside the review can prove helpful [9] in establishing a higher confidence in the accuracy evaluation of a flagged product defect and safety issue. [10] speaks about all the techniques which are essential for pharmacovigilance via social media data in a systematic review.

Few studies on safety related defect discovery stated herewith are equally inspiring. Work by [3] in 2012, which studied online discussion forums resulted in a vehicle defect detection system (VDDS) and the associated framework to assist vehicle product managers with a timely detection of safety issues in Automobiles. In 2015, [11] used the frequencies of occurrence of various smoke words and applied a chi-square feature selection to detect automobile components with safety issues that led to recall of vehicles. The model for predicting the vehicle recalls was trained using supervised ML techniques of Naive Bayes, Decision tree and K-Nearest neighbour classifiers [11]. [12] developed an unsupervised learning based probabilistic defect identification model along with the expectation maximization algorithm to extract multifaceted defect entities which used multinomial distributions to fit the variation in vehicle model, year, component with defect and defect symptom and a Gaussian mixture model distribution for complaint date in the complaint records of National Highway Traffic Safety Administration (NHTSA) in 2016. Another study in 2016 by [13] describes the application of smoke words and text-mining techniques to Toy safety surveillance by analyzing online customer reviews. In 2017, [14] demonstrated the application of domain specific smoke words to analyze joint and muscle pain relief treatments from 3200 Amazon online reviews and compared the performance with traditional sentiment analysis techniques. Smoke words have been used for safety issue detection as shown by [4] for dishwasher appliance in 2017. [15] shows a more generalized and domain independent approach with a Product Defect Latent Dirichlet Allocation (PDLDA) model to identify product defects on three different datasets from Automobile complaint records, problem discussion threads from official forum of Apple for defects in Macbook and extraction of diseases from Patient.info in 2019. In a recent work [16] has put together a thesis depicting the application of safety issue detection in mobile phone products from online consumer reviews using unigram, bigram and trigram smoke terms.

In the remainder of this project, we propose to explore permutations of the techniques of sentiment analysis, smoke terms, machine learning based classification and unsupervised learning based probabilistic approaches to devise an efficient framework to detect safety issues in consumer electronics products.

4 Methodology

4.1 Reviewers and Reviews

Reviewers are in this case consumers themselves. They are categorised as people who have bought a product. When leaving a review, they may not have experienced the product yet. However, for simplification this research assumes that the user had have some sort of experience with the product.

The reviews that are considered are online and public available on many retailer websites. To scope the space of reviews, the work will focus on the worst scored reviews (i.e. 1 out of 5 stars reviews), notice that this information is not given by all retailer platforms. This confinement will help the classifier to learn a more fine granular distinction later on. The reviews are expected to be a mixture of professional, non-professional and informal language. Reviews about not receiving the product or the supplier not responding to messages are not important and need to be distinguished. This is one of the challenges for this task which will be picked up later again.

4.2 Data scraping

As main source Amazon is considered, as it counts to the largest retailer platform world wide and has much free accessible data. The approach described in the following can be equally applied on other retailer platforms like Ebay, Alibaba and Walmarkt for example.

Online communities and platforms like Github steadily provides up to date solutions to scrape data from retailer websites, see Amazon Scraper and Aliexpress Scraper, which can be utilized for this task. However, many of these tools only scrape essential information, i.e. just the review itself. When it comes to trace back and investigate hazardous reviews, more information is needed like a product url or even the store url or manufacturer name. Therefor are more extensive scraping tool has been implemented for this task which is briefly described in the following. All scraping code can be found in the 'Scraping'-folder:

First a product taxonomy from Amazon's electronic product list was created. The distinction of different subcategories will help to investigate and better understand hazards in different product domains. The scraper will create taxonomy of categories on Amazon and their associated url's by scraping trough the menu bar. For electronics this results in about 588 electronic subcategories, each subcategory includes an url which points to the respective area on Amazon. Each subcategory has 2 parent categories (root category, layer1 and layer2). The taxonomy tree for electronics can be found in the in the 'taxonomy.csv' file. There are even more subcategories but which have not been further decomposed.

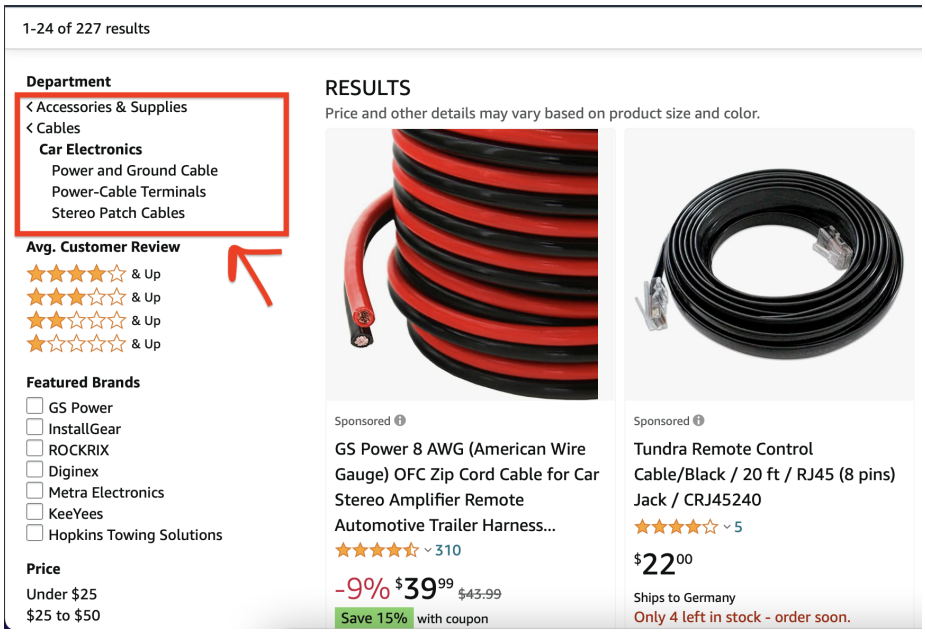

Figure 1: Electronic products categories on Amazon

The taxonomy scraper can also be applied for other product domains on Amazon, i.e. 'Sports & Outdoor'. In the next step the tool will scrape for each of the (588 electronic) categories through a defined number of products and for each product scrape a defined number of reviews. From the obtained products a table will be build with the columns: category (name of product root category, i.e. electronic accessories, cameras etc.), layer1 (subcategory), layer2 (subsubcategory), product title (name of product as it appears in the title), price, brand, manufacturer, manufacturer country, offered since (date of product release), reviews total, rating distribution, reviews collected (number of collected reviews per product), 1-3 star reviews (this column contains the list of reviews together with the star ratings per product), product ASIN (an unique identifier given by Amazon), product url, store name and store url. An example dataset is in the 'Scraping/Dataset'-folder named '5000_electronic_goods_sample_amazon'.

4.3 Data preprocessing

To analyze the reviews several pre-processing steps have been applied to make the reviews easier to compute for deep learning models. This pre-processing steps include simpler steps as removal of white spaces, removal of remaining html and url code in the review, punctuation removal, lowercasing of each word, stopword removal and include more complex steps as lemmatization. Stopwords are frequently words like 'the', 'is', 'will', 'what' which do not necessarily contribute to the task of text classification and which are deleted in this case.

Lemmatization is the process of transforming words to an uniform form, i.e. 'drives', 'drive', 'drove', 'driven' will become 'drive'. In the literature it is advised to use lemmatization over stemming, since it leads to higher accuracy. On the other hand lemmatization needs more computation which might be not feasible in larger datasets. Below is an example for two pre-processed reviews.

```
Gets extremely hot to the point that it would burn my fingers
if I touched it.
=> get extremely hot point will burn fingers touch

Poor quality. Only lasted about 8 months.
=> poor quality lasted 8 months
```

Finally, the review will be tokenized, which means it will represented by a number instead of a word.

4.4 Training Data and Smoke Terms based Labeling

To apply a bidirectional LSTM for text classification, usually large amounts of training data are needed. To tackle this problem a large Amazon reviews database have been used: https://nijianmo.github.io/amazon/index.html. It contains about 21 million electronic goods reviews. From these reviews only the 1-star reviews (about 2.7 million) are considered and others have been discarded.

To be able to train the classifier, the reviews needed to be labeled. Two approaches have been tried for this. First the labeled dataset from the group was used and secondly a unsupervised approach have been tried by filtering candidate reviews based on smoke terms or combination of smoke terms. The filtered reviews will be labeled as positive.
To create a list of smoke terms, smoke words from https://ec.europa.eu/safety-gate-alerts and corresponding synonyms from https://www.synonyms.com could be scraped. In the following are some results from this procedure:

```
['asphyxiation', 'suffocation']
['burns', 'burnes', 'burn', 'scalds', 'combusts']
['chemical', 'chemical substance']
['choking', 'throttling', 'strangulation', 'strangling']
['cuts', 'reductions', 'cutbacks', 'clippings', 'cut-backs', 'cortes',
'lucf', 'cutouts', 'scissions', 'incisions', 'outages', 'compressions',
'haircuts', 'redundancies', 'cut', 'reduces', 'cups', 'reduction',
'decreases', 'denominations', 'pieces', 'savings', 'slashes', 'parings',
'cuttings', 'courts', 'snips', 'declines', 'restrictions']
['damage to hearing']
['damage to sight']
['drowning', 'demersion', 'submersion', 'inundation', 'noyade']
['electric shock', 'shock', 'electrical shock']
['electromagnetic disturbance']
```

```
['energy consumption']
['entrapment', 'trapping', 'impinger', 'quenching', 'trapline', 'sequestration',
'capture', 'trap', 'encapsulation']
['environment', 'surround', 'surroundings', 'environs']
['fire', 'set on fire']
['health risk']
['injuries', 'injury', 'wounds', 'lesions', 'injured', 'wounded', 'casualties',
'accidents', 'trauma', 'traumas', 'damage', 'infections', 'harm', 'injures',
'damages']
['measurement incorrect']
['microbiological', 'microbiologic', 'microbiology', 'microbial', 'micro-organisms']
['security', 'surety', 'protection', 'guarantee', 'safety']
['strangulation', 'choking', 'throttling', 'strangling']
['suffocation', 'asphyxia']
```

This list needed be further processed by hand because some of them should be discarded, i.e. 'haircuts' etc.. The code can be found in the 'Scraping'-folder.

Although for this work more specific terms related to the electronic domain have been used:

```
'smell' and 'burn' (3076)
'cable' and 'melt' (386)
'battery' and 'explod' (383)
'arcing' (97)
'electrical shock' or 'electric shock' (79)
'extremely hot' or 'extreme hot' (1260)
'extremely dangerous' (38)
'storch mark' (28)
'safety alert' (6)
```

All reviews will be labeled positive. In total 5380 filtered positive reviews from 2.7 million 1-star-reviews have been filtered with this method (note the numbers above indicates how much reviews were filtered for each combination of smoke words). The same amount for negative reviews have been sampled randomly.

4.5 Word embeddings

To train the bidirectional classier word embeddings are used. At this state of the project it does not use self trained word embeddings as it used TensorFlows inbuilded embedding layer.

4.6 Model

Neural networks are popular models for NLP tasks and they outperform the more traditional models. Convolutional neural network (CNN) models use convolutional layers and maximum pooling to extract higher-level features. The strength of Long Short Term Memory Networks (LSTM) in text sequences is

to handle long-term dependencies, as they use different units. Although this task is not dependent from the sequence order, the decision was made to use a bidirectional LSTM.

The model architecture consists of three layers, the embedding layer, the bidirectional layer and one dense layer. The bidirectional LSTM receives the embeddings as input and predicts whether it is a positive (safety alert) or a negative review. The maximum input length can be 12 tokens (otherwise it will be cutted) and uses (post) padding for shorter sequences. Is the outcome prediction in the following greater than 0.5 it can be marked as positive (safety alert).

5 Experiments

In this section the model will be evaluated on some self formulated sentences:

```
1:
Input: It already became extremely hot several times!
Preprocessed: already become extremely hot several time
Output: 0.97786075 -> positive (should be positive) = correct
2:
Input: I can recommend this to you.
Preprocessed: recommend
Output: 0.24370304 -> negative (should be negative) = correct
3:
Input: The plastic melted
Preprocessed: plastic melt
Output: 0.76662225 -> positive (should be positive) = correct
4:
Input: The covering melted
Preprocessed: covering melt
Output: 0.18281578 -> negative (should be positive) = incorrect
5:
Input: I smelled fire coming out from my kitchen.
       As it turns out this thing caused it.
Preprocessed: smell fire come kitchen turn thing cause
Output: 0.99969393 -> positive (should be positive) = correct
6:
Input: I saw fire coming out from my kitchen.
       As it turns out this thing caused it.
Preprocessed: see fire come kitchen turn thing cause
Output: 0.00180768 -> negative (should be positive) = incorrect
7:
Input: Warning!! Do not purchase this. I cut my finger with this.
Preprocessed: warning purchase cut finger
Output: 0.27615544 -> negative (should be positive) = incorrect
```

```
8:
Input: The cable melted in the case. I can not recommend to buy this.
Preprocessed: cable melt case recommend buy
Output: 0.02777535 -> negative (should be positive) = incorrect
9:
Input: The cable is extremely robust. I can recommend to buy this.
Preprocessed: cable melt case recommend buy
Output: 0.8398432 -> positive (should be negative) = incorrect
10:
Input: Extremely cool.
Preprocessed: extremely cool
Output: 0.730348 -> positive (should be negative) = incorrect
11:
Input: Each time I plug this I get an electrical shock!
Preprocessed: time plug get electrical shock
Output: 0.16358434 -> negative (should be positive) = incorrect
12:
Input: The battery exploded!
Preprocessed: battery explode
Output: 0.86267513 -> positive (should be positive) = correct
```

As the experiments show there is much improvements for the classifier, as it is trained on too few data to make stronger predictions. Noticeable is that the classifier picks up the filtered words and learns these (see example 1, 5, 9, 10, 11 and 12). Although for some terms as 'electrical shock' (see example 11) it still cannot recognize them. This is because it has to few reviews filtered with the term 'electrical shock' (79) or it was not trained enough (to avoid overfitting). During the experiments it turned out it has learned some new terms, i.e. 'plastic', which seems to appear often in positive filtered reviews (see example 3 and 4). As the data is only trained on 1-star-reviews it might also not being good on predicting positive reviews which contains words from the filtered ones, i.e. 'extremely' (see example 9 and 10). Overall the model predictions are quite vague because it is trained on too few data (less than 10000 after discarding some) which are all truncated to 12 words and repeatedly show the same terms for the positive reviews. For are more fine granular distinction much more reviews which contain hazards would would be needed.

6 Conclusion

The approach for the dataset labeling is naive, because some of the negative labeled samples might contain hazards and some of the positive labeled samples might not necessarily contain hazards. Especially for domains where high accuracy is important this approach might not be valid. On the other hand, if proper smoke terms are chosen, i.e. 'battery' and 'explode', it is more likeli to filter reviews which actually might contain safety alerts. Moreover, this method is very fast (it does not need any human annotators) and helps as starting point

to obtain larger amounts of labeled data.

Deep learning techniques can be facilitated to understand the context of hazardous reviews. As the results show there is slightly context-based learning happening, i.e. terms as 'plastic' which have not been filtered but are learned from the classifier are more likeli to appear within safety alert, because it learns that the term appears in the context of melted cables. As the experiments show, much more data is needed, thus a more fine granular distinction from safety alerts and complaints can be made. That can be tackled through more reviews (which contain hazards), a better data preparation, i.e. by taking the entire review into account (not truncating to 12 words context around the filtered words), by deleting or replacing the crucial (filtered) terms with paraphrases or by applying paraphrasing on the entire review to obtain variations of positive labeled reviews. Moreover, self-trained word embeddings within this domain can lead better results. This work showed a pipeline on how to obtain and pre-process data for the task of hazard detection in consumer reviews. It utilizes modern deep learning techniques as the bidirectional LSTM to enable context-based classification.

References

[1] *Safety Gate: the EU rapid alert system for dangerous non-food products.* URL: https://ec.europa.eu/safety-gate-alerts/screen/search?resetSearch=true. (accessed: 22.02.2022).

[2] Araceli Zavala and Jose Emmanuel Ramirez-Marquez. "Visual analytics for identifying product disruptions and effects via social media". In: *International Journal of Production Economics* 208.C (2019), pp. 544–559. DOI: 10.1016/j.ijpe.2018.12.02. URL: https://ideas.repec.org/a/eee/proeco/v208y2019icp544-559.html.

[3] Alan S. Abrahams, Jian Jiao, and G. Alan Wang. "Vehicle defect discovery from social media". In: *Decision Support Systems* 54.1 (2012). 87, pp. 87–97. ISSN: 0167-9236. DOI: 10.1016/j.dss.2012.04.005. URL: https://doi.org/10.1016/j.dss.2012.04.005.

[4] Darren Law, Richard Gruss, and Alan S. Abrahams. "Automated defect discovery for dishwasher appliances from online consumer reviews". In: *Expert Systems with Applications* 67 (2017), pp. 84–94. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2016.08.069. URL: https://www.sciencedirect.com/science/article/pii/S0957417416304675.

[5] Zhilei Qiao et al. "A Domain Oriented LDA Model for Mining Product Defects from Online Customer Reviews". In: Jan. 2017. DOI: 10.24251/HICSS.2017.222.

[6] Titus Hei Yeung Fong, Shahryar Sarkani, and John Fossaceca. "Auto Defect Detection Using Customer Reviews for Product Recall Insurance Analysis". In: *Frontiers in Applied Mathematics and Statistics* 7 (2021). ISSN: 2297-4687. DOI: 10.3389/fams.2021.632847. URL: https://www.frontiersin.org/article/10.3389/fams.2021.632847.

[7] Lu Zheng, Zhen He, and Shuguang He. "A novel probabilistic graphic model to detect product defects from social media data". In: *Decision Support Systems* 137 (July 2020), p. 113369. DOI: 10.1016/j.dss.2020.113369.

[8] Serkan Ayvaz and Mohammed Shiha. "The Effects of Emoji in Sentiment Analysis". In: *International Journal of Computer and Electrical Engineering* 9 (Jan. 2017), pp. 360–369. DOI: 10.17706/IJCEE.2017.9.1.360-369.

[9] Susan (Sixue) Jia. "Behind the ratings: Text mining of restaurant customers' online reviews." In: *International Journal of Market Research* 60.6 (2018), pp. 561–572. ISSN: 14707853. URL: https://mu.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=133531461&site=ehost-live&scope=site.

[10] Dimitra Pappa and Lampros K. Stergioulas. "Harnessing social media data for pharmacovigilance: a review of current state of the art, challenges and future directions". In: *International Journal of Data Science and Analytics* 8.2 (Sept. 2019), pp. 113–135. ISSN: 2364-4168. DOI: 10.1007/s41060-019-00175-3. URL: https://doi.org/10.1007/s41060-019-00175-3.

[11] Xuan Zhang et al. "Predicting Vehicle Recalls with User-Generated Contents: A Text Mining Approach". In: Lecture Notes in Computer Science : Security and Cryptology 1611-3349. Cham : Springer International Publishing : Springer, 2015, pp. 41–50. DOI: 10.1007/978-3-319-18455-5_3. URL: https://doi.org/10.1007/978-3-319-18455-5_3.

[12] Xuan Zhang et al. "Identifying Product Defects from User Complaints: A Probabilistic Defect Model". In: Aug. 2016.

[13] Matt Winkler et al. "TOY SAFETY SURVEILLANCE FROM ONLINE REVIEWS". en. In: *Decis Support Syst* 90 (June 2016), pp. 23–32.

[14] David Z. Adams, Richard Gruss, and Alan S. Abrahams. "Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews". In: *International Journal of Medical Informatics* 100 (2017). 108, pp. 108–120. ISSN: 1386-5056. DOI: 10.1016/j.ijmedinf.2017.01.005. URL: https://doi.org/10.1016/j.ijmedinf.2017.01.005.

[15] Xuan Zhang et al. "Discovering Product Defects and Solutions from Online User Generated Contents". In: May 2019, pp. 3441–3447. DOI: 10.1145/3308558.3313732.

[16] Muhammad Zeeshan Younas. “Defect Identification for Cell Phones Using Product Reviews”. PhD thesis. 2021.