ISSN: 2754-4745

Journal of Physical Medicine Rehabilitation Studies & Reports

SCIENTIFIC Research and Community

Review Article

Open Access

Customer Feedback Monitoring in E-Commerce with NLP

Keshava Reddy Depa

Manager, Product Management Amazon.com, Seattle, WA, USA

ABSTRACT

The focus of this paper is on applying NLP to enhance the safety monitoring of electronic products. While the online retail market is expanding, product safety plays a critical role in forming customers' trust, preserving the company's reputation, and handling legal concerns. Standard methods of product safety risk management are ineffective in handling such a large number of customer responses, resulting in inefficiency and potential missed risks. NLP appears to provide a possible solution to the proposed problem by analyzing large volumes of customer reviews, comments, and complaint data that most e-commerce platforms receive, allowing them to identify any emergent issues regarding product safety. As the scale of e-commerce continues to grow, monitoring the safety of products becomes increasingly tricky and remains challenging across diverse markets and in different categories of products. It demonstrates how NLP models can be applied to help manage safety feedback, develop robust models against noisy and imbalanced data, and enable multilingual analysis to accommodate global feedback. One of the significant components of the study is the analysis of how NLP can be used in the automation of product safety investigation, which makes this process faster and more accurate. Additionally, the paper looks critically at the future evolution of NLP; it highlights that AI technology and machine learning will enhance product safety monitoring systems more. NLP then promises to help e-commerce platforms identify safemitigateld trust with customers and satisfy legal requirements. The paper concludes by arguing that as NLP technology advances, there will still be more prospects for the development of innovations in product safety that can assist business entities in sustaining their competitive advantage in the sphere of e-commerce.

***Corresponding author**

Keshava Reddy Depa, Manager, Product Management Amazon.com, Seattle, WA, USA.

Received: Janaury 03, 2022; **Accepted:** Janaury 10, 2022; **Published:** Janaury 28, 2022

Keywords: Product Safety, Natural Language Processing (NLP), E-Commerce Platforms, Customer Trust, AI in E-Commerce, Unstructured Data, Product Safety Monitoring, Machine Learning in E-Commerce, Multilingual NLP, Feedback Analysis

Introduction

Determining and guaranteeing the safety of products available in Internet stores is a significant aim for some companies to enter e-commerce platforms, especially if they are oriented to fashion, beauty, and consumer goods. The era of alternative shopping has dramatically influenced e-commerce boost, giving customers more possibilities, but more dangers can appear for the clients while buying something online. While customer safety is often a matter of ethical emphasis for businesses, it is now a fundamental economic concern with profound legal, image, and performance implications. Not understanding or recognizing safety-related issues leads to legal and financial responsibilities, including cases, sanctions, and business legal exposures. In addition, there can be a loss of customers' trust since they will be hesitant to use products from a particular company whose brand they dislike. Hence, speedy detection and management of related product safety hazards is critical for sustaining customers' trust and preserving the image and reputation of a brand.

In the case of large-scale e-commerce platforms, especially those that retail fashion and beauty products, monitoring product safety is a challenging and very demanding process. Most customer returns in e-commerce might be due to size, fit, or quality issues resulting from differences in customer tastes, which are not hard to address. However, there are instances where customers surf narrow nipples to note that they have sharp edges or are broken at the heel and have chemical smells that may harm them or other customers. These are considered Product Safety (PS) cases; these are far more severe and different from the previous cases because of their possible impacts or consequences. To address the PS cases appropriately, there is a need to scrutinize the reported problem to identify if it is a safety issue that deserves additional intervention. The factors considered when diagnosing PS often go through a structured three-workflow process used when searching for cases of PS. Initially, comments from many customers are filtered and evaluated by the FCC agents, which remove or filter out potential PS concerns. Tending to argue that PS relates to actual customers only on comparatively few occasions, the number of receptions mainly contains non-safety complaints. Therefore, out of hundreds, if not thousands, of customer complaints, only a few are taken to the next level of investigation. In the second stage, another PS team is assigned to handle the escalation process, where much attention is paid to scrutinizing possible safety problems in these reports. The PS team determines the level of risk in each case before isolating real quality problems from the rest, filtering out those that are not genuine safety concerns. Only in the most urgent and legitimate safety threat cases are safety threat cases passed on to the laboratory stage. However, this detailed, sequential approach has several practical drawbacks in operating the method. The first is rooted in the scarcity of genuine PS cases, which makes sense when viewed together with the consideration that the term is broadly used as a synonym for marketing-savvy. Since most reported issues are not genuine safety concerns, there

is a high possibility of using up much precious energy, time, and money in cases that do not warrant attention. This can result in enhanced inefficiency in the PS investigation workflow because not-so-heinous cases are taken to the PS team and unnecessarily complicate workflows. Moreover, while cases pass through these stages, the scrutiny level is added to the operational costs, enhancing the business's costs. Thus, the identification of cases with greater probabilities of actual safety risk in the early stage of investigation is essential to ensure the appropriate allocation of investigative effort.

The unstructured customer feedback includes written comments and optional product images often pasted in huge numbers; through NLP, it is possible to automate and thus optimize the investigation process, especially when identifying areas with potential safety risks. This way, the problem of finding cases of PS is formulated as a simple data-driven problem, and given the availability of sufficient training data, simple supervised text classification models can be used to identify key safety signals from customer feedback. It is the logical continuation of the previous methods used in the diagnostics of genuine PS cases and the reduction of numerous false positives, which, in turn, strengthens the PS team resources and further focuses the team on real top-priority issues.

NLP-based framework aims to address several core challenges faced by e-commerce platforms:

- Class imbalance involves many more non-PS-related cases than PS cases.
- Noisy data where customers may provide erroneous reports.
- Data that is multilingual since customers come from many different locations and prefer to use other languages when reporting their experiences.

This means that using advanced NLP in conjunction with data augmentation and strategic training procedures can boost the model's performance to meet actual production use. This paper will also describe how NLP can be used to drive improvements in the handling of product safety investigations to help improve safety for customers and do so in the most efficient manner, saving costs in the process. Consequently, integrating NLP in the product safety investigation process is an excellent chance to update and expand the prospects of product safety in e-commerce. Integrating automation in e-commerce would improve the accuracy of identifying safety concerns, and from the customer's perspective, such problems could be better addressed sooner.

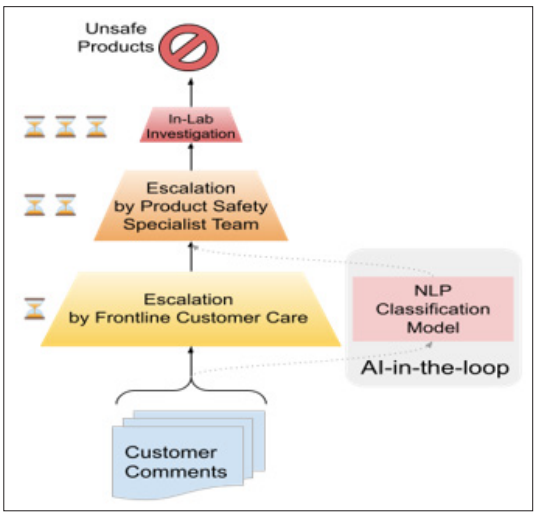


Figure 1: Product Safety Investigation Workflow

The Role of NLP in Improving Product Safety Monitoring

In the dynamic environment of e-business, the safety of products has become one of the most significant concerns in terms of consumer confidence, regulation, and corporate image. It is difficult to identify safety concerns when millions of products in the e-marketplace employ conventional data analysis approaches. Natural Language Processing (NLP) is a method to solve this challenge, which comes under artificial intelligence [1]. Thus, NLP can be a powerful driver for improving product safety supervision by preventing consumer feedback risks, analyzing risk indicators, and indicating them for additional study. The focus of this article is to discuss how NLP could help prevent false claims and better safeguard consumers in the e-commerce business.

The Challenge of Manual Product Safety Monitoring

E-commerce platforms receive massive volumes of general information about customers' actions and opinions daily. Consumers made comments, complaints, and requests containing descriptions of possible risks related to the products. However, such comments are frequently masked and dragged down by thousands of irrelevant comments to safety, making it impossible for a human reviewer to pick out critical issues. There exist many manual stages in the conventional approach to examining PS claims, which are as follows: First, the customers' complaints are some things CS representatives need to sieve through to look for safety risks. Next, product safety teams investigate such complaints and evaluate the degree of problems reported to them. Lastly, some institutional examinations may be referred to the laboratory for analysis. While this workflow is critical in conserving, it is incredibly resource-demanding, time-consuming, and inefficient because of the tremendous amount of information and the relatively few real safety issues at hand.In addition, the overall lack of product safety events, compared to other existing problems, such as sizing disparities or quality problems, results in a marked asymmetry in the data. Most customer complaints are not safety complaints, and this makes the identification of such complaints extremely difficult. In response to these challenges, NLP can be the solution, offering tools for automating the probing of product safety issues for enhanced efficiency and accuracy among business entities [2].

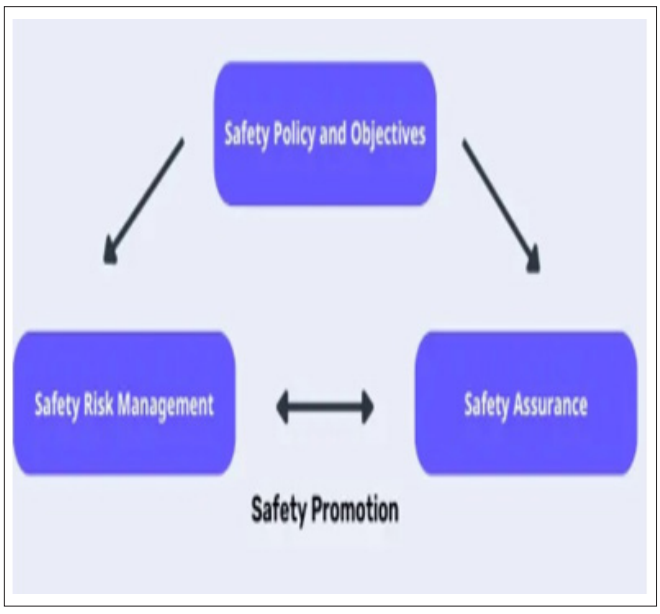


Figure 2: Safety Management System

How NLP Transforms Product Safety Monitoring

Applying NLP can help extract and analyze safety-related signals from extensive unsupervised customer data, substantially enhancing the management of product safety investigations. Here's how NLP makes a difference:

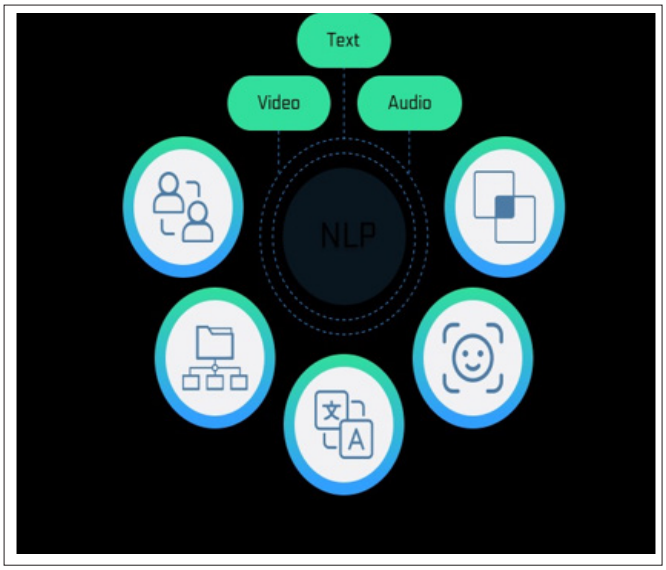


Figure 3: Product Safety Monitoring

Automated Text Classification for Safety Issues

Undoubtedly, the automated text classification procedure is one of the most influential approaches of NLP in monitoring product safety. The breadth of customer comments is distinguishable, whereas NLP models developed by training on labeled data can sort comments into categories such as 'Product Safety,' 'Quality Issue,' and 'Not Relevant.' The patterns of words and phrases that might be associated with product safety risks include sharp edges, toxic fumes, allergies, and all forms of dermal reactions. Since using the text classification for the e-commercial Sites, they will be able to redirect any comments regarding the safety issue to a particular department to attend to it without having to go through all the feedback provided by the customers. Also, the program can work with massive amounts of entries, making it adaptable for vast accounts [3]. It is also helpful for platforms that receive feedback in different languages.

Managing and Under Structure and Novestic Data

In e-commerce, customer feedback is likely to comprise a large amount of unstructured data, which could encompass things like text comments, images, and other forms of input in multimedia format. These significant data sources are unstructured, meaning NLP can be applied to extract information from written commentaries that would otherwise be ignored. NLP techniques can also overcome noisy data – extra, incorrect information that may mislead safety investigators. For instance, natural language processing algorithms must be trained so that during the overall processing, their filter excludes non-safety-related comments that include things like 'Sweat smell' or 'Wrinkled fabric.' Another advantage of using NLP is that it screens out all the trash to only allow worthy causes, especially regarding product safety. As more organizations expand their products and services globally, they require an effective and permanent multilingual capability that can be integrated into the overall structure of their global platforms. Some e-commerce platforms operate internationally, and the feedback received regarding products can be provided in different languages. A particular difficulty that becomes apparent when comparing product safety across various districts is the ability to accurately identify all safety-related complaints, including those reported in languages other than the one in which they were filed. For instance, NLP models such as mBERT and Multilingual BERT can comprehend customer feedback in multiple languages, helping businesses track the safety issues that consumers may face in other countries [4]. Thus, ToRReS-authored NLP systems can help handle input in various languages, therefore providing uniform safety monitoring across regions, which requires less time than relying on the review of English-only feedback may indicate and might overlook potential safety issues in areas where English is not dominant [5].

Real-Time Surveillance and Alarm

In an e-business environment, buyers want quick answers to questions about product safety issues. NLP models can be used in the current monitoring system, which can look at customer feedback as soon as it is posted. If there is anything that could be a possible safety threat, assurances can be made to the concerned groups for subsequent action through alerts provided by the NLP system. NLP in real-time commodity inspection helps address product safety concerns and prevent possible harm to the customer and legal and reputational costs for e-commerce platforms (He et al 2020). Further, it does not create a pile of safety claims that may arise in businesses and harm customers since it attends to these concerns as soon as they occur.

Analyses of Data to Support Organizational Development

The final benefit of using NLP in product safety monitoring is its versatility for analyzing customer feedback. To identify patterns and trends about what customers deem safe in their products, NLP models process and discover repeated patterns and common dangers of product quality defects and faulty material utilized in the production line. Such findings can easily dictate future products and guide the quality checkpoints regarding safety risks before they recur throughout society [6]. The model also has the advantage of customer feedback and can be updated, making product safety monitoring much more efficient as the model incorporates more information.

NLP can be used to shift e-commerce product safety monitoring from a human-based approach to a more rational model based on analysis of consumers' free-form comments to provide for more efficient safety issues identification and resource allocation. Large and complex multilingual data, filtering out noise information and real-time detection of safety-related risks – companies can effectively address customer safety concerns using NLP while minimizing costs and increasing customer confidence. In this regard, increasing the usage of such technologies as advanced NLP in e-commerce processes and, in particular, in the sphere of monitoring product safety is going to be increasingly critical due to the further development of the e-services market [7]. With the help of AI, organizations can be sure that they are not only helping their clients but also supporting brand and organic authority when it comes to shadowy behavior in rather saturated markets.

Problem Description

The use of NLP to analyze and track product safety in e-commerce is also not without challenges, especially regarding the realism of customer reviews and feedback. Three common problems that may emerge the most are class imbalance, noisy data, and multilingual data sets [8]. All these challenges influence the credibility of the NLP models and the speed at which the product safety investigation is conducted

Practical Challenges

Looking at e-commerce platforms, perceptions received always display a massive difference in safety-related issues compared to general quality issues. Whereas customer feedback does not primarily relate to safety issues, that is, size, fit, or general dissatisfaction with the product, product safety complaints constitute a relatively small but significant proportion of total feedback. This is where imbalance is a big problem for any machine learning model, especially for classification where the idea is to separate the binary classes of "Product Safety" (PS) and "Not Product Safety" (NPS). The major challenge of class imbalance is that the machine learning model usually has a prejudice for the majority class, which is the 'Not Product Safety' class in the customer feedback analysis [9]. As the model is trained on far more samples that are not PS-related than samples that are, it tends to memorize and learn how non-PS complaints look like but possibly overlook or misclassify actual PS issues. The above leads to high TPR for the majority label but low TNR for the minority label or product safety issues, preventing efficient identification of safety risks. Several strategies to overcome the problem of class imbalance include Data augmentation, resampling, and class weighting. However, these techniques are not free from defects. Oversampling the minority class tends to increase the rate of overfitting, where the model is trained on minority cases. Under-sampling the majority class could be disadvantageous as it minimizes valuable information. Furthermore, in e-commerce contexts, such imbalance is further elaborated by the deficiency of high-quality labeled datasets for PS cases, making training adequate models challenging.

Table 1: Overview of Investigation Cases

Case Type	Criterion
Allergic Reaction	Rashes/Redness, Itching/ Irritation Pimples, Burning Sensation Swollen skin, and Shortness of breath.
Chemical Smell	Fish smell Petrol/gasoline/diesel smell Sulphur/Chlorine smell Strong dye smell.
Injury	Nail sticking out Needle/Pins/Metal object found Sharp/rough edge Slippery shoes, and Broken heel.
Not Product Safety	Perfume smell, Sweat smell Torn label, Signs of wear Cigarette/smoke smell, Blisters Stains, creases, and scratches.

Multilinguality

Most e-commerce platforms belong to the international market and exist in different countries with different languages and cultures. This entails processing customer feedback in tens of languages for such massive platforms. Considering product safety concerns, the multilingual nature of the data poses an extra problem, as the same issue can be reported in different languages, implying that the invention of a model for identifying PS cases in all markets cannot be straightforward [10]. There are two significant issues at the heart of this challenge here. First, what are the challenges of using learning NLP models that are capable of processing feedback from multiple languages when there are diverse languages? For instance, a complaint about a chemical-like smell in a hotel may require a different approach from a similar complaint in a hotel in the Spanish-speaking world because the two are two separate languages. However, both refer to a safety problem. Also, because of the cultural differences, the terms used to convey safety concerns differ, and thus, the model needs to be adapted to better fit regional expression of product concerns. Second, the matter is even more complicated when the data is annotated in multiple languages. Many e-commerce businesses have a large amount of customer feedback in English, for example, or German, but almost no labeled data in other languages. This results in a problem termed "language bias," whereby an automatic translation system can perform extraordinarily in languages with many examples for training but poorly in languages with few examples. For instance, a model trained primarily with German and English inputs would fail to recognize safety issues in Polish or Croatian with fewer labeled samples.

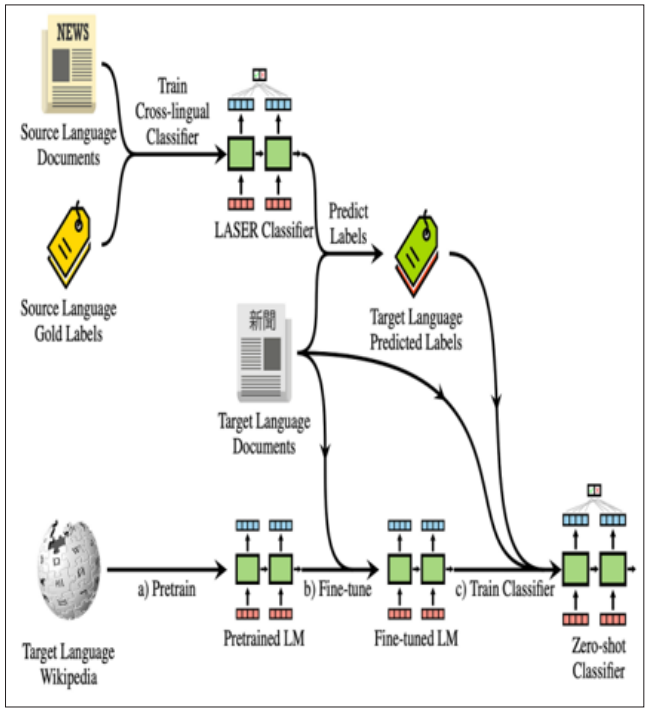


Figure 4: Multilinguity

Most of today's NLP models, including BERT and its multilingual versions, offer the max and latest multilingual BERT (mBERT) or Cross-lingual Language Model Releasing (XLM-R) [11]. However, such models are still limited, particularly concerning the complexity of product safety issues in languages not commonly represented in CAL. One solution is to increase the amount of data by including translations or other parallel texts, which causes problems with data coherence and increases the costs of running the system. Further, getting consistent performance of models for all languages is not easy, even when multilingual models are employed. This makes the model highly effective in languages that have a lot of data but less so in languages that have a few annotated samples, hence the inconsistent performance across markets.

Label Distribution Mismatch

Noisy data can be described as erroneous, irrelevant, or inconsistent data within the dataset. When it comes to noisy data sources in the context of product safety monitoring, it can appear in many forms. Lukewarm customers may not strictly describe a problem with the product and sometimes provide wrong and ambiguous information or categorize a problem as one that affects safety when it does not. For instance, a customer might file a complaint of a broken shoe heel, a quality defect, not a safety defect; they may

notice a strong smell of perfume as a hint of actual or impending toxicity without further proof. Such inconsistencies in the feedback can significantly affect an NLP model's performance level. If the model is trained on noisy data, it may learn incorrect relations to a particular class and thus misclassify even when used in real-world scenarios. For example, if the model is supposed to identify PS issues, it may mistake an issue unrelated to PS, such as a blemish, as a PS issue, hence producing what can be referred to as false positives [12]. On the other hand, the model may overlook potential safety concerns, e.g., products with sharp edges, if the feedback does not describe the seriousness of the problem.

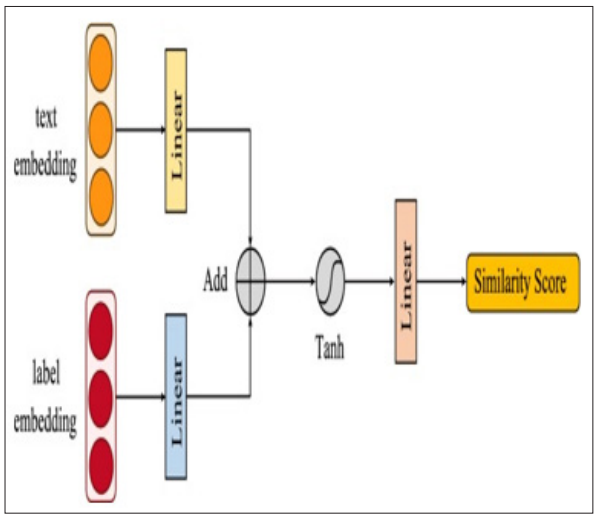


Figure 5: Distribution Mismatch

Mitigating noisy data usually involves cleaning the data set, which consists of identifying error records and their elimination or correction. This can also include excluding comments that contain too many spelling mistakes, comments that are too vague, comments that only contain complaints or gore, or comments that don't match a specific terminology standard across the dataset. While these steps can help to enhance the quality of the data, they do take time and often need either a particular level of manual intervention or are automated only in a complex way. Besides, it can be detrimental to eliminate specific layers of information, especially in safety complaints where other less clear signals may be stripped during cleaning. Additionally, there is an issue with noisy data caused by the variability of descriptions of potential customers' safety concerns.

As much as feedback can be constructive and accurate, it is provided in many forms, including technical writing about product flaws and disgruntled messages. For example, a customer might express that "sharp edges are dangerous" or that they cause "cutting," or might describe an allergic reaction as "rashes" and "skin irritation." Unlike the signal terms, the wording of these phrases is diversified, making it difficult for the NLP model to classify such concerns relating to safety correctly. Large class ratios, noisy data, and multi-language datasets are significant issues impacting performance when using NLP to supervise product safety in online shopping. These issues are relevant regarding developing accurate and reliable models for discovering product safety risk factors in customers' feedback. Solving these issues implies using data augmentation, enhanced pre-processing, and multilingual models. Although efforts are being made, it remains critical to dispel such barriers to achieve high efficiency in NLP-based monitoring systems for product safety, safeguarding customers' interest, and sustaining brand reputation in international markets. E-commerce businesses should further refine ways of handling these issues to better identify real safety concerns in the early stages, operate at a lower cost, and build customer confidence. Eventually, solving these issues with data will become crucial for perfecting and making NLP-based product safety monitoring a dependable instrument for e-commerce.

Problem Formulation

The issue of product safety has become vital as e-commerce platforms solve their area, especially where large amounts of customer feedback are received. Some of the most common manual methods of solving PS issues may result in numerous complaints, most of which are unsafe issues and would prove pricey. The result of intensive studying, as well as the application of Natural Language Processing (NLP), can be a practical and efficient method that will allow nominating and analyzing potential product safety issues in customer reviews and, at the same time, cutting the expenses on operational outgoing and improving the speed of investigations.

Methodology

Adoption of NLP for product safety monitoring, is posed it as a binary text classification issue. Each customer complaint or review is classified into one of two categories: "Product Safety" or "Not Product Safety." This binary approach makes the detection process relatively easier while allowing the model to focus on the types of issues that are not safety-related and which ones should be flagged for safety analysis [13]. Such an approach will still be more effective than a multiclass analysis due to the nature of the problem at hand. In most real-life cases, the members of the product safety team tend to regard all potential safety concerns, including allergies, chemical product smell, and injuries, as genuinely captured cases that need to be investigated further. Since safety issues are not numerous, the goal is to consolidate those situations and prevent them from being examined in terms of time and line and have time for other reactions and non-safety feedback. The binary division of the types of risks also logically excludes possible confusion and contributes to improved performance when distinguishing between what is a threat and what is external to the safety sphere.

Model Architecture

The main constituent of the proposed NLP-based approach is a BERT model, which performs very well in the context of text classification. This technique has benefited from extensive training using datasets and can parse the textual characteristics of customer feedback. Applying BERT to this purpose taps into the advantage of transfer learning in that the model trained on an enormous amount of data can be further fine-tuned, focusing specifically on data regarding product safety. The BERT model of the input text is being tokenized using the shared vocabulary called WordPiece, which converts every customer's review or comment to a set of IDs corresponding to words or subwords. The first string argument to BERT is the tokenized input, and the second one is the 'context-independent word embeddings' that are instead produced by the model BERT after passing the tokenized input through multiple layers of these models [14]. The last representation of the input is due to the [CLS] token to consider it as a summary representation of the whole input sequence. This representation is passed through a linear layer, which projects it onto a two-dimensional vector representing the two output classes. Last, the softmax activation function is used to obtain the likelihood of the modes/numeral classes. The fine-tuning process uses a customer complaint set labeled by the product safety (PS) team [15]. When training, the model uses backpropagation to change its weights to minimize the cross-entropy loss function and accurately classify complaints

under the correct categories. The optimization method used for efficient optimization is Adam, while early stopping has been used to avoid overfitting.

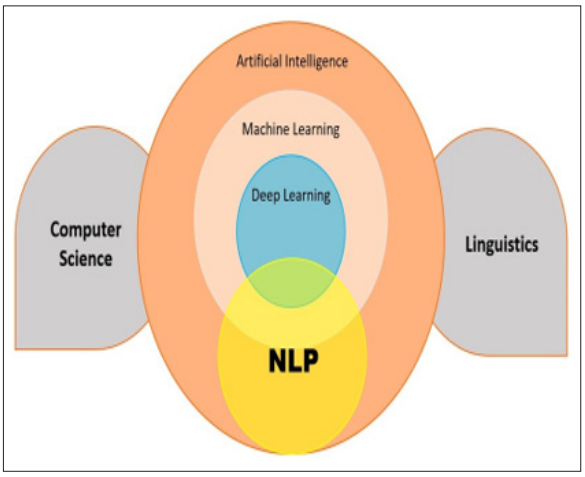


Figure 6: NLP Architecture

Addressing Multilinguality

One significant issue in e-commerce product safety is that the data are multilingual. Many e-commerce platforms are located in multiple locations, and customers provide suggestions in different languages. The first objective of data augmentation is to generate more balanced datasets to reduce overfitting and, hence, enhance the model's capability to accommodate other languages. Additional translations of those customer reviews with annotations extend the model's multilingualism to more languages, such as Spanish, French, and Italian. It also ensures that the model does not have to employ mass translations during the inference process, which might be very slow and costly [16]. Using multiple languages brings forward even more difficulties, including variations in performance across languages in the case of the given model. For instance, the model, trained on highly-accuracy English complaints, loses its efficiency when working with feedback written in other languages that are not frequent in the sample.

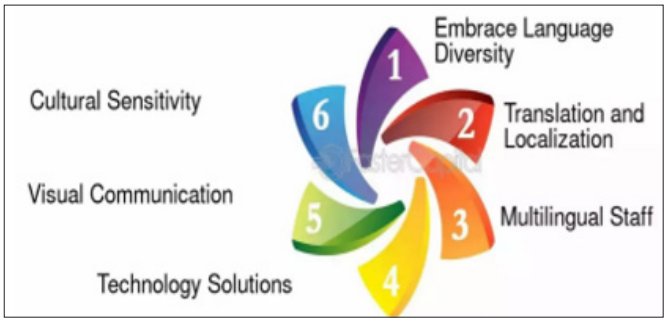


Figure 7: Addressing Multilinguality

Addressing Label Distribution Mismatch

One of the challenges when employing machine learning for accurate data is the reality of the misaligned label distribution between the training and the production environments. Regarding product safety monitoring, the figures for actual safety-related problems are generally small compared to those that are not. This leads to a very skewed dataset, where the number of "Not Product Safety" is significantly higher than the number of "Product Safety." In such cases, the model becomes insensitive to the majority class, creating false negatives, such as missed safety concerns. Data mining technique is employed to overcome this challenge and build a balanced training set. Instead of using methods of under-sampling or over-sampling, more noisy negative samples is mined from the remaining pool of customer feedback the FCC agents have already screened [17]. Such comments are probably not safety issues but are not flagged as non-safety ones simultaneously. BERT model is used to classify these commented phrases, and then to make sure that these are not similar to real safety situations. This approach helps the model select a better decision boundary, reducing the probability of flagging potential safety issues as noise while increasing the speed of recognizing complaints that are not safety risks.

Benefits of the NLP-Based Approach

Increased Efficiency: Applying NLP to CCI improves the analysis time and sorts all complaints, allowing product safety to only work on critical complaints and instances.

Cost Reduction: Specifically, applying NLP in the banking system eliminates the need to employ many workers to attend to small details of a task, hence cutting operational expenses. Finally, due to false positive elimination at an early stage, the system also reduces the number of alerts passing over to the product safety team and laboratory investigations.

Scalability: Using NLP, e-commerce sites can increase their safety measures and cover more languages and geographical locations without a similar increase in cost. The model can always be trained and retrained using new data so it remains dynamic and in touch with the changing concerns of the customer.

Enhanced Accuracy: With the right training, NLP models can provide better accuracy in identifying real product safety concerns compared to keyword-based systems. They can grasp the characteristics and connotations of customer feedback to produce a better classification effect.

Applying the NLP technique to address e-commerce product safety enhances the performance of the checking process, improves the amount of resources, and protects consumers while shopping online with various companies or stores. Incorporating the new generation of machine learning, such as BERT, into these e-commerce sites creates efficient, standard, and safer working environments for their products, increasing customer confidence and satisfaction.

Experiments

As e-commerce platforms grow, proper oversight of products' safety remains paramount at best. Automatically identifying product safety concerns through Natural Language Processing (NLP) models has already been applied, at least in some cases, but its performance should be confirmed [18].

Definition of Data Splits

Precisely, the validity of any NLP model dramatically depends on the ability to assess its efficiency or effectiveness based on the chosen parameters. When evaluating their model, Taigi and Stewart also used precision, recall rate, and F1 rate metrics to analyze the effectiveness of product safety monitoring. Accuracy measures the proportion of the validated noteworthy product safety risk that the model identifies as safety-related [19]. This simply means that if a given model files a high score on the precision level, it does not categorize non-safety complaints as safety complaints. Again, Recall centers on how many actual product safety problems the model could predict. This is particularly important since a model that does not recognize these characteristics (low recall) may fail to identify a risky product and potentially endanger the customer.

The F1 score is the formula for the harmonic mean of precision and recall and is widely used in cases where both approaches must be balanced. The F1 measure of 0.89 is evidence that the model used can identify all the necessary aspects regarding product safety without missing out on minor details [20]. However, a confusion matrix is expected to be used to evaluate the model's performance, where the number of TP, FP, TN, and FN is presented. This means it becomes easier to see that areas where the model is doing well or poorly can be seen in more detail than the whole setting.

Flood and Landslide Models and Experiment Outcome

In the experiments, the NLP model was trained using a large dataset of customer reviews and feedback retrieved from several e-commerce platforms, with a focus on product safety claims. Only the useful information essential for the decision-making process was considered in the form of comments related to the product [21]. After data cleaning and labeling, the model was trained using supervised learning feeds in different stages.

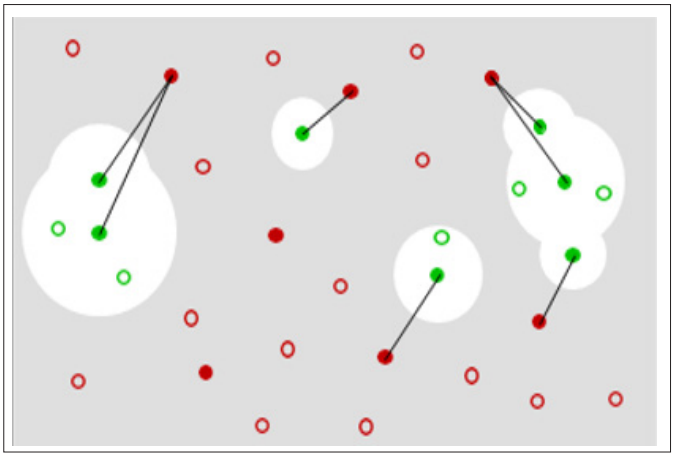
The model achieved promising results across the key performance indicators:

Precision: The model had 0.87 accuracy; therefore, it got 87 percent right; in other words, 87 percent of the claims it produced regarding safety issues were correct. This is very good in its outcome because it reduces the chances of what is referred to as second-generation faults, which are non-safety features identified as belonging to the safety category.

Recall: As it recalled 0.78, the model passed 78% of the actual safety issues. While this presents a good result, there is room for improvement in identifying more of the safety complaints that were not elicited.

F1 Score: This means the model has a good F1 score of 0.82, providing acceptable recall without compromising on precision. This signals the system effectively flags safety issues without flooding the interface with unrelated concerns.

It is important to note that the test set represents a small, labeled subset of the traffic data. To ensure a fair comparison between different models, the test and traffic data is kept consistent across all models [22]. The relevant statistics are summarized in Table 2. The experiments are conducted using three variations of the training data: Original, Original+NN (augmented with Noisy Negatives mined from unlabeled data), and Original+NN+PC [23].



Labeled Positive Labeled Negative Noisy Positive Noisy Negative

Figure 8: Illustration of Noisy Negative Mining

Table 2: Dataset Statistics

Dataset	Train size\#words	Dev size / #words	Test size / #words	Traffic size / #words
Original Original+NN	12.7K / 42.62	1.3K / 35.98		
	281K / 29.20	32.3K / 26.68	11,2K / 32.20	138,3K / 30.22
	418K / 34.60	46K / 29.93		

Evaluation

To fine-tune the model, the changes were experimented to feature extraction and altering the usage of different neural networks. First, the 'bag of words' and conventional techniques (such as logistic regression) were applied. However, this methodology was less successful, as the recall rates were lower and could not address the depth of context necessary in complex safety cases.

The performance improved significantly when BERT is replaced, a deep-learning network. This means that by learning the context within a sentence, BERT noticed safety-related problems, even if the number of other words and phrase constructions overshadowed the latter [24]. The number of recalls and the precision achieved after adopting BERT were higher; thus, this model was used further for continual product safety evaluation.

It is important to note that standard classification metrics such as precision and recall can only be evaluated using the test set. To assess compliance with the second business requirement, the number of comments classified as positive from the total traffic data is reported (referred to as Model ∪ FCC). The BERT models for each of the three datasets using FLAIR was fine-tuned [25]. The hyperparameters used include a mini-batch size of 32, a learning rate of 3e-6, and early stopping based on development set loss. To optimize for the first business requirement, a prediction threshold based on achieving 95% recall on the corresponding dev set for all models was determined [26].

Table 3 indicates that the model trained on the Original dataset, due to class prior mismatch, results in more than 16,000 cases being forwarded in a month. The model trained with the Original+NN dataset performs better, reducing the forwarded cases to 3,400 due to the inclusion of Noisy Negatives. The proposed model, trained on the Original+NN+PC dataset, achieves the highest recall of 92% on the test data while also minimizing the number of forwarded cases to 2,700.

Table 3: Experimental Results with Different Training Methodologies

Model Training Dataset	Precision Recall	Volume	Model ∪ FCC	Volume Model	Avg. Std.
Original	0.60	0.78	16, 165	16, 102	0.20
Original+NN	0.75	0.52	3416	3342	0.15
Original+NN+PC	0.63	0.92	2782	2714	0.12

The proposed training method achieves the highest Recall, and lowest Volume of escalated comments.

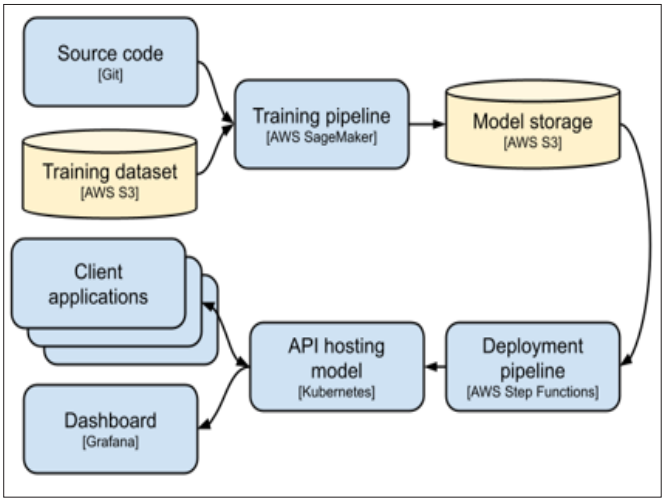


Figure 3: Production Deployment of the Model

Language Fairness: One issue one faces when using an NLP model for product safety monitoring on international e-commerce platforms is ensuring the model is fair across various languages and markets [27]. Given the large variety of customer reviews, it is critical to have a probability distribution over the languages that do not put some languages ahead of others, or the model will see only a portion of the safety-related reviews adequately. For instance, customer complaints in languages lacking training materials may be misinterpreted, and thus, safety signals are missed [28]. On the other hand, the model may excel in notoriously large languages like English because of too many data inputs but might not be effective in markets that require linguistic and cultural differences. In response to this, a multilingual approach to NLP was used. The model was created based on English reviews and Spanish, French, and Chinese data. As data was augmented in such a manner, the model was made more robust, with better results realized with the model in different areas.

After implementing multilingual training, the following results were observed:

Fairness Evaluation: The model produced similar levels of accuracy in different languages, though it is important to note that the datasets represented only a selected range of languages. The precision slightly declined for languages with more elaborate grammatical structures, such as Chinese. However, the recall rate did not differ significantly across languages, indicating that the model was just as effective at identifying incidents that might cause safety as existed across the language barrier [29].

Language-Specific Challenges: A few languages needed refinements based on specific contexts. For instance, the informal phrases and expressions used by Spanish speakers were previously categorized as unrelated to safety. It was later noted that more data sources in Spanish languages were included in the training.

In the rightmost column (Avg. Std.), how consistently each model handles comments across different languages was presented. For every comment in the test dataset, the translation versions in other languages was included. Ideally, a model should produce identical likelihood scores for a comment regardless of its language, resulting in a standard deviation of 0 across different language versions. The average of such standard deviations for all samples in the test set was calculated, with lower scores indicating better performance. Despite starting with a multilingual BERT model, the model trained on the Original dataset has the highest average standard deviation. In contrast, the proposed Original+NN+PC model achieves the lowest standard deviation, demonstrating that training with a parallel corpus is necessary for fair treatment across languages.

Production

These applications of NLP models towards bilateral communication are well scalable, and as more firms adopt e-commerce, coupled with the explosion of feedback received daily, the subsequent possibility of product risks increases, hence the necessity to adapt NLP models to operate in large-scale user environments for businesses interested in customer safety. While building these models in a controlled environment is critical, deploying them into production triggers new issues about infrastructure, performance, and improvement [30]. This section covers how NLP models are deployed into production, how performance is then monitored, and what issues companies continue grappling with regarding model effectiveness.

Training and Deployment of NLP Models: Using Pipelines

It is essential to develop a solid pipeline for training and deploying NLP models to achieve these goals. The pipeline presented here helps to maintain the continuity of model deployment from the development phase into production, and here's how it benefits e-commerce platforms:

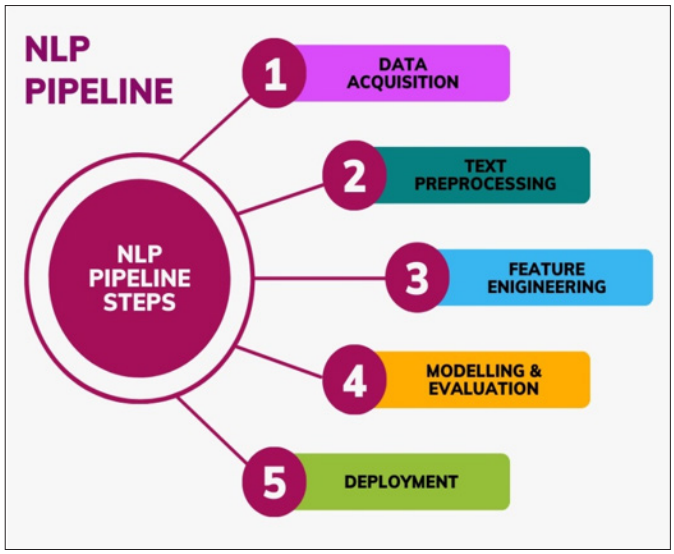


Figure 10: NLP Pipeline

Taking the AWS Sage Maker for Model Training

AWS SageMaker is one of the most widely used cloud-based services for machine learning, and it can be used to design, train, and deploy machine learning models at scale [31]. It provides a highly autonomous solution where all the intricacies of model deployment, such as servers' creation, models' updating, and servers' scaling, are solved automatically. For NLP models focused on product safety, SageMaker provides several key advantages:

Automatic Scaling: SageMaker adjusts the model's infrastructure in response to the demand for resource utilization to process large volumes of e-commerce feedback in real-time.

Training with Large Datasets: The platform is also suitable for distributed training, which allows for handling overwhelming customer feedback across different languages and products.

Managed Pipelines: It makes AWS SageMaker Pipelines contain the features that can enable the full cycle of machine learning automation, from data capturing to model checking [32]. This will help minimize human interaction and possible errors in the deployment stage. For the same, SageMaker also offers pre-trained NLP models that are BERT or GPT, on which fine-tuning may be done for e-commerce platforms for a particular purpose, such as product safety detection. At the same time, this characteristic makes it possible to implement updated models swiftly, guaranteeing the system's adaptability to trends and customer preferences.

Serving the Model: Kubernetes and Git

Even if a model is developed, using it in a production environment requires planning to avoid instability. This is usually accomplished using Kubernetes, open-source software for deploying containerized applications. It helps businesses conveniently use NLP models where multiple containers are handled across various clusters.

Kubernetes Offers Several Benefits for NLP Model Deployment: Containerization: When NLP is deployed in containers, the system guarantees that the model is the same for the various environments, such as a business's development, staging, and production phases, without compromising the dependencies.

Scalability: Kubernetes can dynamically increase the range of containers needed, accommodating the model's need to address more customer feedback during busy periods such as the holiday season or the release of new products.

High Availability: Another aspect that Kubernetes implements is high availability, which is built into placing the load, thereby decreasing the possibility of a system failure [33]. Also, it used a Version Control System like Git for the model code. In this way, it easily defined different versions by updating configuration and dependencies for models and rolling back changes. This is especially important when working with multiple versions of the model for A/B testing or fine-tuning the model based on customer feedback. The purpose of writing this paper is to understand performance monitoring when it is applied in realistic scenarios. The move of NLP models into production brings new considerations into how you should and can monitor the performance of these models. E-commerce businesses need to monitor closely how well the model picks out product safety risks and how effectively it addresses live feedback from customers. Lack of monitoring can cause potential safety problems that cannot be spotted and thus affect the levels of customer satisfaction, trigger legal sanctions, or tarnish the company's image.

Checking on How the Models are Evolving

Once the NLP model is trained and deployed, it needs to be closely monitored to ensure the results generated are accurate and timely. Critical metrics for monitoring NLP models in production include:

Precision and Recall: Accuracy tests the efficiency of optimistic prediction or, in this case, correct safety issues, while recall evaluates the number of actual problems that the model recognizes. Regarding product safety, these two measures are mutually essential and must be achieved in that proportion. High recall guarantees that most of the safety issues will be identified, while high precision eliminates the labeling of many instances as unsafe.

Response Time: Secure and real-time monitoring is crucial in e-commerce businesses. This means that NLP models should be able to quickly analyze large amounts of customer feedback and identify the safety problems that need to be solved [34]. Throughput: Throughput can be defined as the number of items, for example, transactions or feedback that the model can handle at a particular time. High throughput allows models to accommodate this feedback in real-time without the throughput limiting the models. To do this, e-commerce applications apply performance monitoring software, including Prometheus and Grafana. These tools are used for metrics, a real-time dashboard for tracking punitive changes, and alert status generation. These tools assist organizations in detecting any signs of performance decline or developing problems that may affect customers.

Feedback and Model Update Tools Real-Time

In an ever-changing e-commerce environment, product safety issues may shift quite often. For instance, new ladders of unsafe products may appear, while other problems might be observed to be more frequent than before. Hence, it is essential to have tools to toggle the coefficients and monitor the model with feedback that can be updated as quickly as an experiment.

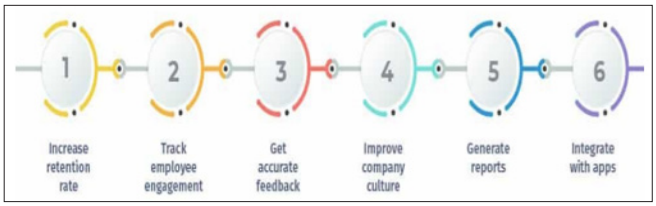


Figure 11: Realtime Feedback Tool

Active Learning: Several learning strategies allow the model to clarify and seek human help in defining some inputs as uncertainty [35]. For instance, when the model tracks a possible safety problem but doesn't have all the necessary data to decide, it sends it to the next level. This way, human input learning allows the model to fine-tune and improve.

Model Retraining: The model can be retrained at specific intervals using real-time performance and newly labeled data, such as customers' complaints or new hazardous product reports. AWS SageMaker of Kubeflow offers ways to do this to prevent the updates from being prompted by humans. Some difficulties concerning the Production and its constant advancement Novelties and opportunities in the process of production

Despite the many advantages of NLP for product safety monitoring, there are several challenges when deploying these models at scale:

Data Drift: The frequency of customer feedback distribution may shift over time because of changes in products or services, the market, and the customers. Data drift is also a concept that makes a given model not perform as expected. This has to be done, emanating from the rate at which the changes are taking place, making it essential to conduct frequent model evaluations and retraining.

False Positives and Negatives: To the authors' knowledge, false positives and negatives are inevitably introduced whenever NLP models and their corresponding applications are used at a large scale. Offsetting these errors is vital to sustaining model efficiency and customer satisfaction. For improvement continuity, businesses should establish a feedback mechanism that captures customer grievances, products being recalled from the market, or any other safety issues so that the model can still be improved.

NLP models used in e-commerce on product safety can be considered both an advantage and a disadvantage. Companies can scale and manage their NLP models through AWS SageMaker and containers and Kubernetes containers for container orchestration [36]. For this reason, checks, constant supervisory feedback, and occasional model recalibration are conducted to ensure the models' efficacy in adapting to the current conditions of the market. However, solving such issues as data drift and false positives is a never-ending process, which means that work in this field will never stop and becomes one of the main components of any NLP solution.

Conclusion

Integrating NLP in the surveillance of product safety in e-commerce applies a new knowledge area in improving consumer protection, operation effectiveness, and brand confidence. This article provided an understanding of the role of product safety in online retail, the exploits of e-commerce platforms, and the revolutionary factor of NLP. As the number of customer reviews, complaints, and questions grows, the attempts to track the safety of products using conventional techniques only are ineffective and time-consuming. NLP solves the general difficulties outlined above as a tool because it implies fast and accurate analysis of vast amounts of unstructured data. It also enhances the speed and accuracy of identifying hazards. The application of NLP in implementing product safety monitoring has several benefits, as analyzed below. Moreover, NLP helps analyze customer feedback faster and ensures that potential problems connected with safety can be prevented before they turn into tragedy. There are often specific keywords associated with a particular product that would indicate that it is unsafe for consumption; binary classifications, model fine-tuning, and multilingual support techniques can then be implemented to detect where these products are accurately and in which markets and languages. Also, when customers' opinions are regularly monitored, electronic commerce venues can address the risks that relate to product safety before they pose serious problems that negatively affect consumers and brands.

There is a lot more that NLP can be applied to in e-commerce that has not yet been explored. Future developments in AI, particularly in NLP, further refinements of the Safety Management safety management systems, and even superior models, including transformer-based architectures and deep learning techniques, will likely enhance the detection of product safety even further. Besides, as many NLP algorithms are developed to analyze the context and distinguish subtle differences in customer feedback, they further differentiate between actual safety issues and heard complaints. Furthermore, the prognosis for the global development of e-commerce is still the future expansion of NLP to increase the safety of products for consumers speaking different languages and living in various countries. Hence, proficiency in performing NLP on multilingual data will remain critical for developing NLP models to improve companies' product safety monitoring worldwide. Since the dawn of the digital economy, as companies venture out to new territories, it will be paramount that their NLP models are well-equipped to address linguistic variations and other cultural differences to protect the fairness and accuracy of product safety evaluations. NLP in the e-commerce domain will also create new prospects for firms to embed themselves in the market more uniquely as time progresses. In the same way, AI safety monitoring will help a company enhance customer satisfaction and cement a loyal and trusting customer base. As product safety gains significance in the market and consumers' preferences, applying state-of-the-art NLP techniques will play a decisive role in e-shopping platforms to compete in the market and safeguard consumers. Therefore, using NLP to improve product safety monitoring is destined to become one of the standards in today's e-commerce systems. The future of AI and machine learning shows signs of enhancing more precise, faster, and customizable approaches to protect consumers and improve the shopping experience. This way, with the constantly developing spectrum of e-commerce, those companies will be ready to implement those advances and set the pace for both the safety of the offered products and the trust of the customers [37].

References

1. Chowdhary K, Chowdhary KR (2020) Natural language processing. Fundamentals of artificial intelligence 603-649.
2. Vajjala S, Majumder B, Gupta A, Surana H (2020) Practical natural language processing: a comprehensive guide to building real-world NLP systems. O'Reilly Media https://www.oreilly.com/library/view/practical-natural-language/9781492054047/.
3. Dickey G, Blanke S, Seaton L (2019) Machine learning in auditing. The CPA Journal 89: 16-21.
4. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency 610-623.
5. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization, in ICLR (Poster)
6. Voigt P, Von dem Bussche A (2017) The EU General Data Protection Regulation (GDPR): A Practical Guide, 1st ed., Cham: Springer International Publishing https://www.scirp.org/reference/referencespapers?referenceid=2996831.
7. Roblek V, Meško M, Bach MP, Thorpe O, Šprajc P (2020) The interaction between internet, sustainable development, and emergence of society 5.0. Data 5: 80.
8. Song G, Huang D, Xiao Z (2021) A study of multilingual toxic text detection approaches under imbalanced sample distribution. Information 12: 205.
9. Borrellas P, Unceta I (2021) The challenges of machine learning and their economic implications. Entropy 23: 275.
10. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, et al. (2021) Ethical and social risks of harm from language models https://arxiv.org/abs/2112.04359.
11. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 4171-4186.
12. Aziz F, Haq AU, Ahmad S, Mahmoud Y, Jalal M, et al. (2020) A novel convolutional neural network-based approach for fault classification in photovoltaic arrays. IEEE Access 8: 41889-41904.
13. Xu Z, Saleh JH (2021) Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. Reliability Engineering & System Safety 211: 107530.
14. Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics https://www.scirp.org/reference/referencespapers?referenceid=3422225.
15. Dhina MM, Sumathi S (2021) An innovative approach to classify hierarchical remarks with multi-class using BERT and customized naïve bayes classifier. International Journal of Engineering, Science and Technology 13: 32-45.
16. Romero F, Li Q, Yadwadkar NJ, Kozyrakis C (2021)

{INFaaS}: Automated model-less inference serving. In 2021 USENIX Annual Technical Conference (USENIX ATC 21) 397-411.

17. Kaghazgaran P (2020) Crowd and AI Powered Manipulation: Characterization and Detection Doctoral dissertation, Texas A&M University
18. Galli F (2021) Algorithmic business and EU law on fair trading https://orbilu.uni.lu/handle/10993/50697.
19. Gomez M, Weiss M, Krishnamurthy P (2019) Improving liquidity in secondary spectrum markets: Virtualizing spectrum for fungibility. IEEE Transactions on Cognitive Communications and Networking 5: 252-266.
20. Wu X, Zheng W, Xia X, Lo D (2021) Data quality matters: A case study on data label correctness for security bug report prediction. IEEE Transactions on Software Engineering 48: 2541-2556.
21. Gursoy D (2019) A critical review of determinants of information search behavior and utilization of online reviews in decision making process invited paper for luminaries special issue of International Journal of Hospitality Management. International Journal of Hospitality Management 76: 53-60.
22. Satheeshkumar K, Raja JT, Kirubakaran C, Anbuselvan B (2021) An effective text mining approach for product safety surveillance using latent semantic Dirichlet allocation, Materials Today: Proceedings https://www.researchgate.net/publication/349383698_An_effective_text_mining_approach_for_product_safety_surveillance_using_Latent_semantic_Dirichlet_Allocation.
23. Taunk K, De S, Verma S, Swetapadma A (2019) A brief review of nearest neighbor algorithm for learning and classification. In 2019 international conference on intelligent computing and control systems ICCS 1255-1260.
24. Tasioulas J (2019) First steps towards an ethics of robots and artificial intelligence https://www.jpe.ox.ac.uk/wp-content/uploads/2019/07/Tasioulas-1.pdf.
25. Akbik AT, Bergmann D, Blythe K, Rasul S, Schweter R, et al. (2019) Flair: An easy-to-use framework for state-of-the-art NLP, in NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) 54-59.
26. Arar ÖF, Ayan K (2015) Software defect prediction using cost-sensitive neural network. Applied Soft Computing 33: 263-277.
27. Ullrich C (2019) New approach meets new economy: Enforcing EU product safety in e-commerce, Maastricht Journal of European and Comparative Law 558-584.
28. 28. Sletvold H, Nguyen T (2021) Experiences and perceptions of foreign-language customers on medication information received in the pharmacy–a focus group study. International Journal of Pharmacy Practice 29: 330-335.
29. Leopold H, Eid-Sabbagh RH, Mendling J, Azevedo LG, Baiao FA (2013) Detection of naming convention violations in process models for different languages. Decision Support Systems 56: 310-325.
30. Leng J, Zhang H, Yan D, Liu Q, Chen X, et al. (2019) Digital twin-driven manufacturing cyber-physical system for parallel controlling of smart workshop. Journal of ambient intelligence and humanized computing 10: 1155-1166.
31. Mishra A (2019) Machine learning in the AWS cloud: Add intelligence to applications with Amazon Sagemaker and Amazon Rekognition. John Wiley & Sons https://onlinelibrary.wiley.com/doi/book/10.1002/9781119556749.
32. Talele GC (2016) What Are The Key Areas Of ML-Ops/DL-Ops In Business Problems For Company Growth Using Cloud Environment?. Global journal of Business and Integral Security https://www.gbis.ch/index.php/gbis/article/view/396.
33. Vayghan LA, Saied MA, Toeroe M, Khendek F (2019) Kubernetes as an availability manager for microservice applications https://arxiv.org/abs/1901.04946.
34. Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: From big data to big impact. MIS quarterly 1165-1188.
35. Hattie JA, Donoghue GM (2016) Learning strategies: A synthesis and conceptual model. npj Science of Learning 1: 1-13.
36. Bhattacharjee A (2020) Algorithms and Techniques for Automated Deployment and Efficient Management of Large-Scale Distributed Data Analytics Services (Doctoral dissertation, Vanderbilt University) https://irbe.library.vanderbilt.edu/server/api/core/bitstreams/900e9ea7-53f9-4caa-a015-170dac91ec65/content.
37. He L, Han D, Zhou X, Qu Z (2020) The voice of drug consumers: online textual review analysis using structural topic model. International Journal of Environmental Research and Public Health 17: 36-48.