Auto Defect Detection Using Customer Reviews for Product Recall Insurance Analysis

Titus Hei Yeung Fong*, Shahryar Sarkani and John Fossaceca

Department of Engineering Management and Systems Engineering, The George Washington University, Washington, DC, United States

The challenge for Product Recall Insurance companies and their policyholders to manually explore their customer product's defects from online customer reviews (OCR) delays product risk analysis and product recall recovery processes. In today's product life cycle, product recall events happen almost every day and there is no practical method to automatically transfer the massive amount of valuable online customer reviews, such as defect information, performance issue, and serviceability feedback, to the Product Recall Insurance team as well as their policyholders' engineers to analyze the product risk and evaluate their premium. This lack of early risk analysis and defect detection mechanism often increases the risks of a product recall and cost of claims for both the insurers and policyholder, potentially causing billions of dollars in economic loss, liability resulting from the bodily injury, and loss of company credibility. This research explores two different kinds of Recurrent Neural Network (RNN) models and one Latent Dirichlet Allocation (LDA) topic model to extract product defect information from OCRs. This research also proposes a novel approach, combined with RNN and LDA models, to provide the insurers and the policyholders with an early view of product defects. The proposed approach first employs the RNN models for sentiment analysis on customer reviews to identify negative reviews and reviews that mention product defects, then applies the LDA model to retrieve a summary of key defect insight words from these reviews. Results of this research show that both the insurers and the policyholders can discover early signs of potential defects and opportunities for improvement when using this novel approach on eight of the bestselling Amazon home furnishing products. This combined approach can locate the keywords of these products' defects and issues that customers mentioned the most in their OCRs, which allows the insurers and the policyholders to take required mitigation actions earlier, proactively stop the diffusion of the detective products, and hence lower the cost of claim and premium.

Keywords: product recall insurance, machine learning, artificia lintelligence, natural language processing, neural network, opinion mining, product defect discovery, topic model

INTRODUCTION

In the Product Recall Insurance landscape, as the number of product recall events increases every year in almost every industry, insurance companies are constantly faced with the difficulty of predicting recall risk, assessing the potential of loss, and estimating the premium. While manufacturers have been using advanced quality control tools for product development, defective products can still be found on the market, and product recall events often happen [1]. The motivation is this paper is to develop a novel approach, combined with RNN and LDA models, to provide the insurers and the manufacturers with an early view of product defects.

These product recall events can cost both the insurance companies and the manufacturers billions of dollars of loss and bring significant brand reputation impact that lasts for an extended period. A good example is the recent 2016 Samsung Galaxy Note 7 explosions recall event. This event cost Samsung more than $5 billion of loss and the subsequent loss of sales in the electronics industry [2]. Allianz Global Corporate & Specialty (AGCS) has analyzed the average value of each product recall insurance claim, excluding small value claims, to be about 1.4 million Euro between 2012 and 2017 [3]. Product recall insurance is intended to provide manufacturers with financial protection for the cost of the recall and their liabilities but when the number of cases increases the insurers have become the biggest victims.With the growth under the new era of Web 2.0, various social media and e-commerce sites such as Amazon.com and Twitter.com now provide online virtual communities for consumers to share their feedback on different products and services. Such feedback, which always includes customer complaints and defect information about the products, can provide valuable intelligence, such as the product's design, performance, and serviceability, for the insurers and the manufacturers to take remedial actions to avoid potential recalls. While there is a large amount of such information available on these sites, there is a lack of an effective method to automatically distill the information to the engineering team.

In the Samsung Galaxy Note 7 recall case, a study showed that there were early customer reports of overheating problems through online customer feedback before Samsung realized and took action [4–6]. It is, however, a challenge for the insurers and the manufacturers to manually read through a large amount of customer feedback available online and be able to discover early signs of product defects, delaying the manufacturers to take necessary recovery actions.

As technology progresses with faster computers and better computational algorithms, we are now able to collect, process, and extract useful information from a large amount of textual customer feedback. In the field of business and product design, studies [7, 8] show that machine learning predictive models can successfully extract useful information from OCRs, including customer buying patterns and customer requirements on future product design. This research provides the foundation for applying machine learning predictive models to detect defective product information from OCRs.

This research explores methods of extracting product defect information from online customer reviews (OCRs) and demonstrates a novel predictive model using Recurrent Neural Network (RNN) and Latent Dirichlet Allocation (LDA) topic model. The new predictive model provides the insurance companies and the manufacturers with an early view into product defects, enabling them to make an assessment of product recall risks, and take required mitigation actions early to proactively stop the spread of the defective products. This will help both sides to prevent further damages and economic losses.

The scope of this research studies the viability of a probabilistic model with RNN and LDA to analyze, extract, and identify defective product information from OCRs. RNN is a type of neural network model for analyzing time-series data. This model can solve problems involving sequences of word order textual data. LDA is a type of generative probabilistic model for discovering a set of topics best describes a collection of discrete data. This consists of the collection of customer online review data with manual labels for supervised learning, quantitative models for testing the hypothesizes associated with identifying product defects, and verification of models.

The new proposed approach includes two quantitative RNN classifiers and one LDA topic model. The first RNN classifier differentiates negative OCRs from non-negative OCRs. The second RNN classifier differentiates OCRs that contain defect information from OCRs that do not contain defect information. The LDA topic model, which combines the first and the second RNN classifiers, generates a set of key product defect topics for a product using OCRs.

The input data consists of 9,000 randomly selected OCRs, and their star ratings, from the furniture section of Amazon's Customer Review Public Dataset. This customer review dataset will be used as input for the RNN classifiers.

The rest of the paper is organized as follows: in *Data*, we discuss related literature reviews. Then, in *Methods* and *Results*, we present a detailed explanation of the data and our proposed model. After that, in *Conclusion* we present the results. Lastly, in *Discussion and Future Work*, and *Data Availability Statement*, we end with conclusion, discussion and future work, respectively.

RELATED WORK

Product Defect and Product Recalls

These defective product recall events can cost manufacturers as well as their insurers billions of dollars and bring significant impact that lasts for an extended period. The United States Consumer Product Safety Commission (CPSC) has concluded that deaths, injuries, and property damages from consumer product incidents cost the nation more than $1 trillion annually [9]. According to AGCS, these events have caused insured losses of over $2bn over the past five years, making them the largest generator of liability losses. An empirical, event-time analysis found that product recall events had a direct impact on the company's equity price for two months after the events' release [10]. These kinds of events expose companies not only to economic damages, but also negative consequences such as loss of goodwill, loss in product liability suits, and loss to their rivals [11].

Due to increased product complexity and more stringent product safety legislation [12], studies show that the trend of product recall events is increasing and no industry is immune from a product recall event [13]. Recalls happen when the manufacturer does not address the issue or was unaware of it before the product was distributed to the market. The main two reasons for recalls are 1) a consequence of design flaws, which make the product fail to meet required safety standards, and 2) manufacture defects in products that do not conform to specifications such as poor craftsmanship [14]. According to Beamish [15], in the toy manufacturing industry, design flaws contributed to 70.8% of the recalls, while only 12.2% of the recalls were from manufacture.

Time to Recall and Product Recall Strategy

The strategy and the timing for the recall of a defective product have a direct impact on both the finances and the reputation of manufacturers and insurers. Time is an essential factor during a defective product recall event. The longer the defective product remains in the marketing and distribution process, the harder it is for the company to take recovery action, and poses more potential for injuries [16, 17]. Studies showed that manufacturers with a preventive recall strategy in place, such as continuously identifying product defects and initiating voluntary recalls, have a shorter time to recall than companies with a reactive approach such as initiating recall after a hazard is reported [17]. There is also research proving that a positive customer and public relations impact on the company will result when a proactive recall strategy is implemented during a product defect event [18, 19]. While it is hard to avoid the existence of defective product risk, the aforementioned studies have shown that time to recall and a company's recall strategy have a direct impact on the company's reputation and losses, which supports the fact that when a company has an early view of product defect, it can reduce the loss of business and damage to reputation, and regain consumer trust.

Opinion Mining With Online Customer Reviews

With the rise of Web 2.0, social media and customer review sites have enabled companies to discover consumer feedback on their products with increased speed and accuracy. Information embedded in CORs has a direct impact on companies and their products [20]. Comparing to "Offline" word of mouth customer opinions, OCRs have a much more significant impact because of their persistent, easily accessible, and open-to-public format [21]. Companies have been looking at OCRs to improve their product and marketing strategies [22]. OCRs enable companies to monitor customer concerns and complaints, as well as to take corrective actions [20]. Some companies even respond to these customer text reviews personally to improve their service [23].

Product Defects Discovery

While there are several research studies and proposed algorithms for using the previous generation products' OCRs to provide valuable information to engineers on the next generation of product development and product design, there is little attention in academia for using OCRs in the later stages of the product cycle to discover customer complaints and product defects [24]. Abrahams et al. proposed a new algorithm using a sentiment analysis method to classify the type of product defect information (e.g., performance defects, safety defects, non-defect, etc.) embedded in OCRs for vehicles [24]. They recognized that while traditional sentiment analysis methods can successfully identify complaints in other industries, they fail to distinguish defects from non-defects and safety from performance defects in the automotive industry. This is because OCRs in the auto industry that mentions safety defects have more positive words, and fewer negative words and subjective expressions than other OCRs. Alternatively, their team spent 11 weeks building and tagging a set of automotive "smoke" words dataset from the OCRs before doing sentiment analysis. This method has shown success in defect discovery and classification, but it is also highly domain-specific for the automotive industry. Bleaney et al. studied and compared the performance of various classifiers, including Logistic Regression, k-nearest neighbors (k-NN), Support-Vector Machines (SVM), Naïve Bayes (NB) classifier, and Random Forest (Decision Tree) on identifying safety issues ("Mentions Safety Issue," "Does Not Mention Safety Issue") embedded in OCRs in the baby product industry [6]. They found that the Logistic Regression classifier had the highest precision, with 66% in the top 50 reviews surfaced. Zhang et al. proposed an unsupervised learning algorithm using the LDA topic model's method to group and identify key information and words in each type of defect from complaints and negative reviews [25].

While these studies have shown some success in using sentiment analysis methods to extract defective information from OCRs, these studies have not been able to locate defectives with OCRs from a product level or accept all OCRs from a single product.

Natural Language Processing

In the field of linguistics and computer science, there is active and ongoing research on how to improve how computers understand human behavior and language. The development of Natural Language Processing (NLP), which is a type of artificial intelligence concentrated on understanding and manipulating human language, has achieved practical successes over the years. NLP models have successfully helped researchers solve real-world human text processing problems, especially research on using OCRs to extract product defects [6, 24, 26]. In this research, Latent Dirichlet Allocation (LDA) and Neural Network and Recurrent Neural Network (RNN) are used to extract product defect information.

Text Representation

Word embedding is a widely adopted method in representing raw textual data to its low-dimension property. This encoding scheme transforms each word into a set of meaningful, real-valued vectors [27]. Instead of randomly assigning vectors to words, Mikolov et al. have created GloVe, one of the most

widely-used pre-trained datasets among researchers for mapping words to vectors, which is to be used in this research [28]. This dataset maps words that have closer English meaning to a closer vector scale in a general sentiment analysis task. An example would be mapping the term "king" in a closer scale to "man" in vector, while further from "woman." Other studies have also built models on top of these two datasets to enhance word embedding in domain-specific sentiment analysis tasks [27, 29, 30].

Latent Dirichlet Allocation

Another unsupervised NLP method that can discover textual data insight is the Latent Dirichlet Allocation (LDA). LDA is a three-level hierarchical Bayesian model that can discover a set of unobserved groups, or topics, that best describe a large collection of observed discrete data. As Blei et al. suggested, the goal of LDA is to "find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks, such as classification, novelty detection, summarization, and similarity and relevance judgments." [31]. LDA was first proposed and used for discovering population genetics structure in the field of bioengineering in 2000 [32] and further used in NLP processing on textual data in 2003 [31].

Researchers have been using LDA to discover topics from a large collection of OCRs to help the industry to gain insight into their product. Santosha et al. and Zhai et al. both suggested using LDA for grouping and producing an effective summary of product features from a large collection of OCRs. Santosha et al. successfully used the product features terms from the LDA topic model to build a Feature Ontology Tree for showing product features relationships [33], while Zhai et al. built a semi-supervised LDA with additional probability constraints to show product features linkage between products [34]. Researchers have also suggested LDA topic model can be used on serving industries to discover business insight such as a summary of OCRs on travel and hospitality review sites. Titov et al. and Calheiros et al. both have suggested using LDA outputted topic's terms to discover and analyze customer reviewer's sentiment for businesses to improve customer experience [35, 36]. In the field of product defect management, Zhang et al. used the LDA topic model to discover short summaries of product defects and solutions on a large amount of online product negative reviews and complaints to help engineers and customers discover product defect information [25].

Neural Network and Recurrent Neural Network

A neural network is a set of connected computational input/output units that loosely model the biological brain [37]. A neural network can be trained iteratively with supervised data, can recognize or "Learn" specific patterns embedded in this data, and perform prediction tasks based on these learned patterns without pre-programmed rules. In the context of this research, the neural network plays an important role in solving text classification problems, especially when using recurrent neural network architecture [38]. Based upon the Neural Network

architecture, researchers in the 1980s have been proposing adding recurrent connections between nodes to solve problems involving sequential data, which is now called the Recurrent Neural Network (RNN) [39, 40]. While this type of network can solve sequential recognition, Bengio et al. found that it is difficult to solve problems where the sequences are getting longer and prediction depends on input presented at an earlier time [41]. This is due to vanishing gradients where the error gradient propagating back tends to vanish in time [42]. Hochreiter et al. also proposed the Long Short-Term Memory (LSTM) approach, a particular type of RNN architecture, which further improved the problem involving long data sequences. LSTM overcomes this problem by adding gates in RNN nodes to regulate the flow of information [42]. Researchers suggested that LSTM can greatly improve the accuracy of sequence learning, such as offline handwriting recognition [43], as well as text classification problems that involve word order [38]. An LSTM-RNN model can build the relations between sentences in semantic meaning on text classification, which can increase the model accuracy over that of the traditional methods [44, 45].

DATA

Data Collection and Data Labeling

The primary data source of this study is the OCRs and their associated metadata from the Amazom.com marketplace. This dataset contains customer text reviews, product information, star ratings, review dates, and other relevant information. The customer reviews dataset will be used as input for sentiment analysis to identify negative customer reviews, determine their usefulness in providing information about defects, classify the types of defects found in the OCRs, and extract product defect topics within the context of this research.

Amazon's Customer Reviews Public Dataset is an organized version of OCR data in a number of tab-separated values (TSV) files for researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML). This data set contains more than 130 million OCRs and associated metadata in 43 product categories in the United States marketplace from 1995 until 2015.

Furniture Customer Reviews and Data Format

The furniture section subset of this dataset was used in this study. This data subset contains 792,114 OCRs and associated metadata of customer options on Amazon furniture products. Due to the constraint of manual labor in creating a supervised dataset, this study randomly selected 3,000 Amazon OCRs with a rating of three or more stars and 6,000 Amazon OCRs with a rating of one or two stars for model building, training, and testing. A sample data of this dataset is shown in **Figure 1**. The focus of this study is the textual analysis in the "review_body" column of the dataset.

Data Labeling

Successful machine learning models are built on large volumes of high-quality training data. To build RNN models that can extract defective information embedded in each OCR, each OCR needs

```
{    "marketplace":       "US",
     "customer_id":       "52703681",
     "review_id":         "RZ0Y9U30658TB",
     "product_id"         "B00000ITPY",
     "product_parent":    "676928760",
     "product_title":     "Edutiles Foam Playmats Uppercase Alphabet 26 Pieces - Complete Alphabet Set",
     "product_category":  "Furniture",
     "star_rating":       "5",
     "helpful_votes"      "5",
     "total_votes":       "6",
     "vine":              "N",
     "verified_purchase": "N",
     "review_headline":   "Educational AND fun!",
     "review_body":       "My son and I BOTH loves these!  This toy is extremely versatile.  One minute they are a floor
mat, the next minute a puzzle, the next he makes them into blocks to stack or he's lining up the letters and
practicing the  alphabet.  Somewhat expensive, but worth the investment.",
     "review_date":       "2000-03-17"}
```

FIGURE 1 | Simple amazon raw data.

TABLE 1 | Definition of AWS SageMaker ground truth labels.

Label	Definitions of Labels	Example
Manufacturing defect	OCRs that mention products that have manufacturing defects and/or are improperly manufactured with physical parts, apart from its intended design	I Was very happy with the purchase for the first 3 or 4 days. Then the bearing dropped out of the tube in the middle of it. (review_id: R238K8EITCNRZZ
Problematic design or quality	OCRs that mention products that have a problematic design or quality issue, with no mention of physical parts falling off	The knee cushion is not comfortable for sitting any longer than about 10 min. The chair is clunky and hard to move on the floor without picking it up. (review_id: R31BYJESH8F2DO)
Bad customer service	OCRs related to the frustration of delivery of the product or customer support process, while not related to the physical products themselves	I Have since called twice more with no returned call. This is the worst customer service I have ever received (review_id: R3VG1CFNR60ED)
No defect information provided	OCRs did not mention any of the defect information from the last three categories	We bought four of these, they are just some real cheap chairs that are overpriced. (review_id: R2F8RCR0LFI7SS)

Instructions ✕

View full instructions

View tool guide

1. **Manufacturing defect**: Customer Reviews that mention products which have manufacturing defects and/or are improperly manufactured with physical parts, apart from its intended design. ***Ex: "I was very happy with the purchase for the first 3 or 4 days. Then the bearing dropped out of the tube in the middle of

This task is to classify the type of defect in a furniture product mentioned in Amazon customer text reviews. This is a machine learning research project conducted by a doctoral student from George Washington University.
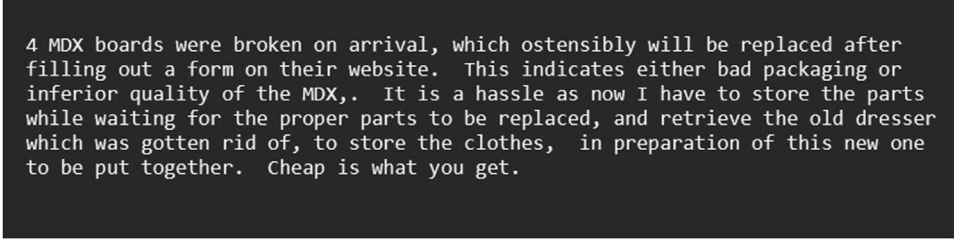
"DO NOT BUY THIS ITEM!!!!! I received this item in broken condition to begin with. The largest snowman of the group was not even close to 14 inches tall. It is made of very cheap materials and is NOT meant for outdoor use. I would not pay any more than $5 for this item in perfect condition. I bought this item from National Tree in October and I am just now contacting Amazon, in January, to finally get my refund. You have been warned."

Select an option

Manufacturing defect	1
Problematic design or quality	2
Bad customer service	3
No defect information provided	4

Submit

FIGURE 2 | AWS SageMaker interface for labelers.

FIGURE 3 | Simple OCR before text preprocessing step.

to be manually labeled by human labelers for RNN model training and testing purposes. The "Amazon Web Services (AWS) SageMaker Ground Truth" data labeling service was procured to manually label OCRs to build the required supervised dataset. AWS SageMaker Ground Truth provides a platform for independent labelers to label machine learning tasks, and each of the 9,000 OCRs was reviewed by three human labelers to ensure the accuracy of the data. Human-labeled results are also generated with a confidence score for each label to ensure high-quality data.

Labelers were provided with a detailed definition of each label with an example (as shown in **Table 1**) to categorize each OCR on the AWS SageMaker Ground Truth interface in **Figure 2**.

Each of the OCRs was labeled with one of the following four labels: "Manufacturing defect," "Problematic design or quality," "Bad customer service," and "No defect information." The labels of "Manufacturing defect" and "Problematic design or quality" are identified as the two main reasons for a company to initiate product recalls [14, 15]. "Bad customer service" is also added as one of the labels because a customer service issue is also where customers often report issues, especially while shopping on an online platform. The human label is able to label and determine which label is the OCRs related to the most in the minimalist and clean interface as shown in **Figure 2**.

Data Preprocessing

Data preprocessing is an essential step in turning raw text data from OCRs into useful information that the computer can read, and machine learning can process. After the OCRs have been extracted from the review body, the OCRs were cleaned and preprocessed, and then turned into digitized text representation vectors.

Text Preprocessing and Cleaning

Human-created text data often contains inconsistent wording, special characters, and contractions, which can contribute to inaccurate data analysis and affect model performance [37, 44, 46]. For this reason, text preprocessing and cleaning is an essential step in ensuring input data quality with normalizing words and removing unnecessary characters. The following three steps are taken in this study to improve text data quality: 1) Stemming, 2) Stop-word removal, and 3) Special character, numeric, and empty text removal.

Stemming is an NLP technique to groups and reduces different words with the same root and linguistic meaning into the same word stem or root form. This study employed the Python Natural Language Toolkit (NLTK) package algorithms for the stemming process. The computer would treat these different words with the same word stem as an equal text vector representation. Stop-word removal can remove over commonly used words in English that give no or very little linguistic meaning to the overall context of the given text. To avoid losing the textual message in translation, this study used a custom-written Stop-word removal function. Special character, numeric, and empty text removal is an NLP technique to remove the non-text characters to improve data quality such as "!", "@", "#", and empty space. This increases the machine learning model's performance by only focusing on real textual data.

The following two figures, **Figures 3, 4** show an OCR before and after these three steps were done.

The two figures, **Figures 3, 4** show an OCR before and after these three steps were done. The special character, empty text and stop-words are removed to improve data quality. Stemming process is also done in the text. Words such as "replaced" are replaced with the root "replac" and "arrived" are replaced with "arriv". This allows the models later in the process treat words with the same root with the same meaning.

Text Representation

As mentioned in *Text Representation*, OCRs textual data have to be turned into text representation in a digitalized format inputting to machine learning models for computers to recognize the information. This encoding scheme transforms each word into a set of meaningful, real-valued vectors to represent each word in a given text [27] This study uses Word2Vector pre-trained datasets for mapping real number vectors to words with similar linguistic meaning to a closer vector scale [28]. This increases the efficiency of the training process for the machine learning model. The data preprocess also takes a text padding step to normalize the length of the ORCs by padding "0" after the end of each text representation before the OCRs are fed to the machine learning model.

```
mdx board broken arriv ostens replac fill form websit indic either bad packag
inferior qualiti mdx hassl store part wait proper part replac retriev old
dresser gotten rid of store cloth prepar new one put togeth cheap get
```

FIGURE 4 | Simple OCR after text preprocessing step.
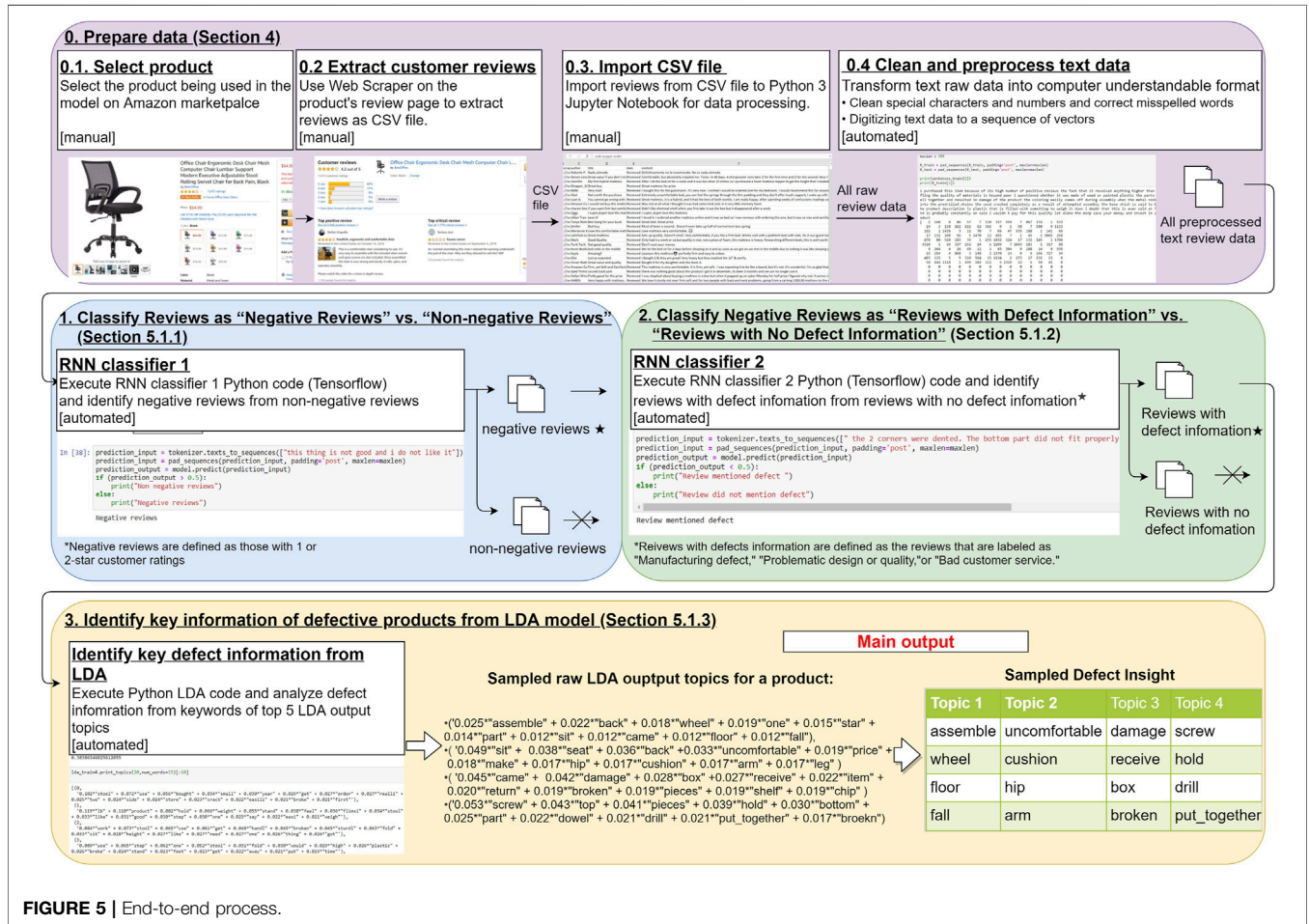

FIGURE 5 | End-to-end process.

METHODS

Model Development and Testing Procedures

The goal of this study is to develop and evaluate how the recurrent neural network (RNN) classifiers and the Latent Dirichlet Allocation (LDA) topic model can extract product defect information from online customer reviews (OCRs). After preprocessing the data, there were three targeted statistical models used in the research in order to accomplish the thesis statement of providing engineers with an early view of product defects. They include two quantitative RNN classifiers and one Latent Dirichlet Allocation (LDA) topic model. The first RNN classifier categorizes negative OCRs from non-negative OCRs. The second RNN classifier categorizes OCRs that mention defects from those that do not. The LDA topic model

provides engineers a view on groups of word items or topics that best describe the type of defects found in a single product.

Figure 5, shows the end-to-end process as to how OCRs from a single product is processed to extract product defect insight using the fully trained and built models. The user of the model would first select the specific product and do data preparation on the dataset as described in *Methods*. After the data is prepared, the data would then go though the three models as described in the following passage in *Results*.

Model 1: Recurrent Neural Network Classifier for Classifying Negative Revies From Non-Negative Reviews

OCRs with negative sentiment has a much higher chance to include complaints and defective issues about products as
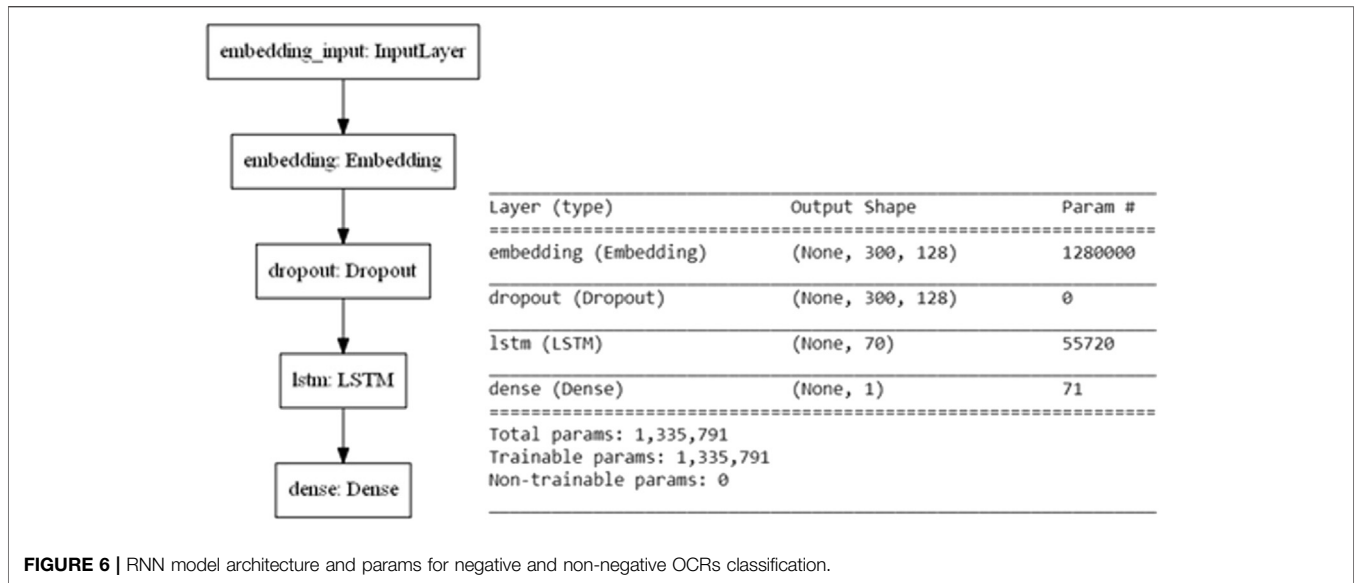
FIGURE 6 | RNN model architecture and params for negative and non-negative OCRs classification.
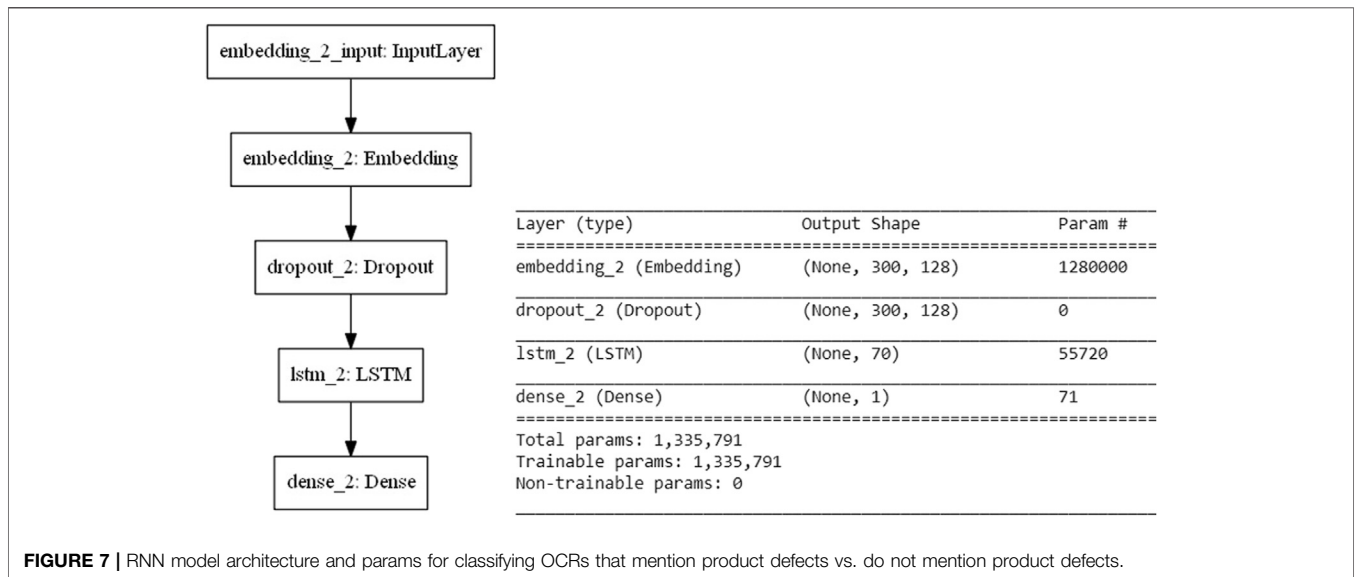

FIGURE 7 | RNN model architecture and params for classifying OCRs that mention product defects vs. do not mention product defects.

compared to OCRs with a non-negative sentiment. A classifier built to distinguish negative OCRs from non-negative OCRs is useful for giving engineers insight into the evidence of a product defect. This RNN model contains five layers, as shown in **Figure 6**.

The first layer is the input layer that takes input vectors. The second layer is for word embedding, as was described in *Text Representation*. The third layer is a dropout layer. The dropout layer randomly sets the neuron's output to 0 during each iteration of training to avoid overfitting. The fourth layer is an LSTM layer with 128 LSTM neurons chained in sequence for recurrently processing information during the training step. After the information is processed with the chained 128 LSTM units, information is

then sent to the output layer for classification. The output layer uses one Sigmoid to separate the output into two classes, either near 1 or near 0. This identifies the OCRs as either negative OCRs or non-negative OCRs. This model uses the Adaptive Moment Estimation (Adam) optimizer for backpropagation to update the hidden LSTM layer with a learning rate of 0.01. The whole dataset it consists of 9,000 OCRs with one-third non-negative OCRs and two-thirds negative OCRs.

The model is then trained with 8,100 OCRs, which is 90% of the total selected data with 50 epochs and a batch size of 32 OCRs in each batch. A successfully trained RNN model is able to automatically identify negative and non-negative OCRs with no given label. This allows engineers to identify negative

OCRs from online forums or markets that only contain text reviews but no ratings. To test the accuracy of the RNN model on classifying negative OCRs from non-negative OCRs, a set of 900 OCRs are separated for testing, which is 10% of the total selected data, with one-third non-negative OCRs and two-thirds negative OCRs. The RNN would then predict the labels of the testing data and compare it against the actual label.

Model 2: Recurrent Neural Network Classifier for Classifying Reviews That Mention Product Defects From Reviews That do not Mention Product Defects

To further investigate which negative OCRs contain product defect or issue information, a second RNN classifier is built to identify OCRs that provide defective product information from OCRs that do not. This RNN model is similar to the first one that contains five layers, as shown in **Figure 7**.

Some negative OCRs do not contain defective product information. For example, review id #R2F8RCR0LFI7SS states: "We bought four of these, they are just some real cheap chairs that are overpriced." This OCR only mentions the product is overpriced with no other information about the product issue. The second example is review id #R3W0KKHC5LK7K5 with the review title as "One Star," and a two-word review in the text body as "Pure junk." These do not give engineers any information about the product itself. The RNN model is trained with the two labels: "Defect information provided" and "No defect information provided". The OCRs that are manually labeled as "Manufacturing defect," "Problematic design or quality," or "Bad customer service" are considered with defective product information provided. Otherwise, OCRs with human labels of "No defect information" is considered as no defect information provided.

A successfully-trained RNN model is able to automatically identify OCRs that provide defect information from OCRs that provide no defect information without a given label. This allows engineers to identify and filter out OCRs that are of value for their engineering design and product correction. To test the accuracy of the RNN model on classifying OCRs that mention product defects from OCRs that do not mention product defects, a set of 900 negative OCRs, which is 10% of the total, was used for testing. The RNN would then predict the labels of the testing data and compare it against the actual label.

Model 3: Latent Dirichlet Allocation Topic Model for Providing Product Defect Insight

The Latent Dirichlet Allocation (LDA) topic model is able to automatically offer engineers a view on groups of words, items, or topics that best describe OCRs with defect information. The LDA topic model builds a probabilistic text model by viewing a document, or OCR, as a mixture of topics, each with its distribution of topics [47]. This allows manufacturers to have an early view of where the problem is.

The LDA topic model assumes that OCRs are represented as random mixtures over k latent topics, where each topic is characterized by a distribution over words [31] After a large number of the iterative training process of reassigning words to topics according to the multinomial distribution, the model is converged with the K numbers of topics each with word distribution that best describe the set of OCRs. The words that make up the word distribution are able to tell the information about the topics among the group of selected OCRs. This allows engineers to have an early view on the OCRs that provided defective information, before manually looking at each one of them.

While LDA models are able to give a view of a list of topics that a group of OCRs is mentioning, for its unsupervised learning property, there is no actual label for the model to test against. The most common evaluation metric that is used on LDA models is topic coherence, which measures semantic similarity among the top words that appear in the word distribution for a single topic. This method might not include the actual meaning of the words that tie back to the defective information. Chang et al. suggested these traditional topic coherence metrics are negatively correlated with the measures of topic quality developed, and they agree that human judgments and manual determinations remain a better way to determine if the LDA model is giving out informative topics among a set of documents or OCRs [48].

To test out this LDA approach along with RNN models that filter out OCRs that mention defective product information, a set of OCRs of the eight top-selling home and furniture products on Amazon were selected for testing. This process developed eight test cases. The OCRs from each product were first filtered using the first RNN model for selecting negative OCRs and then filtered using the second RNN model for selecting OCRs that mention defective product information. These OCRs were then inputted into LDA for discovering essential topic information on defective product information. The 10 highest-scoring words that build upon the five LDA topics were used for testing purposes. Amazon SageMaker labelers were then asked to identify if these topics, or the group of words, are related back to the defectiveness of a product or the details of a product itself. This method verifies whether the output topic words are linked back to the product or the defect itself.

RESULTS

This section provides comprehensive results of the three presented models conducted during this study to give engineers an early view of product defects and issues. The first two models are recurrent neural network (RNN) classifiers that categorize OCRs that contain product defect information. This section also demonstrates the results of these models against the human labels of both the model training and the model testing stages. The last model is the Latent Dirichlet Allocation (LDA) topic model that can extract product defect information from OCRs. A total of eight test cases for the LDA model using the eight bestselling products on the Amazon.com home furniture section was used. This section presents the output of the LDA model along with its relevance to the defective product information.

TABLE 2 | Prediction statistics and hypothesis test of for RNN model classifier for classifying negative OCRs from non-negative OCRs.

Prediction Statistics

True Positives	509	Testing accuracy	0.8500
False Positives	48	Testing precision	0.9156
True Negative	256	Testing F1 score	0.8851
False Negative	87	Testing recall	0.8593
		One sample Z-tests for a proportion	
		H_0: $p = 0.7$	
		H_1: $p > 0.7$	
Sample proportion	0.850000	95% confidence interval for proportion	(0.826672, 0.873328)
Z-value	9,82	p-value	0.000
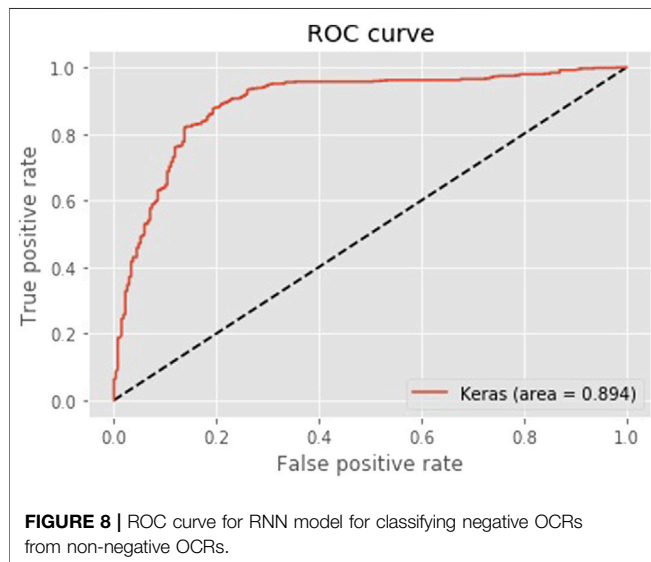
FIGURE 8 | ROC curve for RNN model for classifying negative OCRs from non-negative OCRs.

Model 1: Recurrent Neural Network Classifier for Classifying Negative Reviews From Non-Negative Reviews

The first model is the RNN classifier for sorting negative OCRs from non-negative OCRs. Negative OCRs are defined as those OCRs with 1 or 2-star customer ratings. Non-negative OCRs are those defined as OCRs with three or more stars customer ratings. The model was trained with 8,100 OCRs and was validated with 900 OCRs, with one-third non-negative OCRs and two-thirds negative OCRs. The 8,100 OCRs were the training dataset reserved for training the RNN model. The 900 OCRs, which had never been trained by the model, were the testing dataset. The model was executed for 50 epochs, with a batch size of 32 OCRs in each batch, at a learning rate of 0.01.

This test evaluates the performance of the RNN Model for classifying negative reviews from non-negative reviews. The test data that was not trained by the model was predicted after the 50 epochs compared with the output against the actual label. **Table 2** shows the prediction metrics and hypothesis test of this model. This table shows the hypothesis test of the one sample Z-tests for a proportion at alpha 0.05 with 70% accuracy. **Figure 8** shows the ROC curve of this model.

The model correctly predicted 765 OCRs out of 900 OCRs. This was tested with a one-sample Z-test for a proportion to evaluate if the model has an accuracy of 70% accuracy. With the area under a ROC Curve at 0.930, the model has good predictive power. With the p-value at 0, the hypothesis test successfully rejected the null hypothesis, and thus this RNN classifier is sufficient in classifying negative reviews from non-negative reviews.

Model 2: Recurrent Neural Network Classifier for Classifying Reviews that Mention Product Defects From Reviews that do not Mention Product Defects

The second model is the RNN classifier for categorizing OCRs that mention product defects from OCRs that do not mention product defects. OCRs that mention product defects are defined as the OCRs that are manually labeled as "Manufacturing defect," "Problematic design or quality," or "Bad customer service." OCRs that do not mention product defects are defined as the OCRs that are manually labeled as "No defect information". This model was trained with 8,100 OCRs and was validated with 900 OCRs. The training data is the data used for training and fitting to the RNN model. The testing data is the data used for testing the model, which was not adjusted by the model. This model was executed for 50 epochs, with a batch size of 32 OCRs in each batch, at a learning rate of 0.01.

This test evaluates the performance of the RNN classifier for classifying OCRs that mention product defects from OCRs that do not mention product defects. The test data that was not fitted by the model was predicted after the 50 epochs and results were compared against the actual label. **Table 3** shows the prediction metrics and hypothesis test of this model. This table shows the hypothesis test of the one sample Z-tests for a proportion at alpha 0.05 with 70% accuracy. **Figure 9** shows the ROC curve of this model.

The model correctly predicted 779 OCRs out of 900 OCRs. The predictive power of this model is good with the area under a ROC Curve at 0.894. The hypothesis test was tested with one-sample Z-tests for a proportion to evaluate if the model has an accuracy of 70%. With the p-value at 0, the hypothesis test successfully rejected the null hypothesis, and thus this RNN classifier is sufficient in classifying reviews that mention product defects from reviews that do not mention product defects.

TABLE 3 | Prediction metrics and hypothesis test of RNN model for classifying OCRs that mention product defects from OCRs that do not mention product defects.

Prediction Statistics

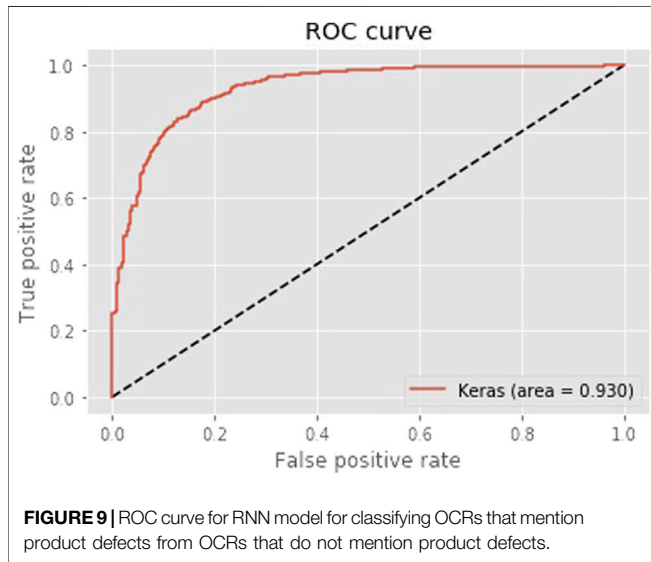|			
True positives	579	Testing accuracy	0.8659
False positives	69	Testing precision	0.8985
True negative	200	Testing F1 score	0.9208
False negative	52	Testing recall	0.9208
		One sample Z-tests for a proportion	
		H_0: $p = 0.7$	
		H_1: $p > 0.7$	
Sample proportion	0.865556	95% confidence interval for proportion	(0.843269, 0.887842)
Z-value	10.84	p-value	0.000

FIGURE 9 | ROC curve for RNN model for classifying OCRs that mention product defects from OCRs that do not mention product defects.

Model 3: Latent Dirichlet Allocation Topic Model for Providing Product Defect Insight

The LDA topic model was used in this study to identify groups of words, or topics, that best describe OCRs that have already been identified as negative and embedded with defect information by previous models. The selection provides the aggregated LDA topic modeling result of the top eight furniture products on Amazon.com as of 12th April 2020. Each product was provided with five topics with 10 words each. Those groups of words were validated by Amazon SageMaker human labelers to verify their relevance to the product. Amazon SageMaker human labelers were asked to evaluate if at least half of the group of words retrieved by the LDA topics are relevant to detail, buying process, usage, or defect of a furniture product.

Example Test Product: Linenspa 2 Inch Gel Infused Memory Foam Mattress Topper, Twin

This is an example of one of the eights that were used in this research. Linenspa 2 Inch Gel Infused Memory Foam Mattress Topper is a polyurethane memory foam mattress topper that adds softness to the top of mattresses to enhance the sleeping experience. This product received 12,684-star ratings at an average of 4.7 out of five stars and 7,202 valid OCRs. The RNN classifier for classifying negative OCRs from non-negative OCRs identified 1,658 negative OCRs among the total valid OCRs. The second RNN classifier for classifying reviews that mention product defects from reviews that do not mention product defects identified 1,363 OCRs that mention product defects.

The LDA topic model was used to identify the five topics with 10 words each that best describe the OCRs which mention product defects and generated a topic Coherence Value (CV) of 0.4930. The following **Table 4** shows the topics, their supporting words, and their support weight. The topics were also given to Amazon SageMaker human labelers for their relevance to a furniture product or a product defect, along with its corresponding confidence level.

Amazon SageMaker human judgment indicated five out of five topics show words related to furniture products or product issues. The five topics summarized the terms used in OCRs that mention defects. Words that support topic number five might indicate a body support problem with the topper being too soft and the foam sinks. The following list shows three sampled OCRs that indicate body support problems using the word search function with the word "support" in the data.

> "Six months later the foam is squashed and offers no support at all. She weighs less than 110 lbs. I will not buy this pad again." (review ID # R2MV2APVP8P9CX)
>
> "This topper goes completely flat once you lay on it and offers zero support." (review ID # RYTH07FT8W7XS)
>
> "This thing provided no support, it crushes down to nothing anywhere there is the weight (I'm 5′9″ 155 lbs … Shouldn't be an issue)" (review ID # RRPFIQ56HYMVU)

A total of eight test product was evaluated with a one-sample sign test in this research with a total of 32,301 ORCs ran through this model. Amazon SageMaker human labelers were asked to evaluate if at least half of the group of words retrieved by the LDA topics are relevant to detail, buying process, usage, or defect of a furniture product. Since there are eight test cases demonstrated in this paper, one sample sign test was used for the hypothesis test. **Table 5** shows the number of relevant topics generated by each of

TABLE 4 | Topic model words for LINENSPA Mattress topper.

Topic	Words supporting the topic (support weight)						Is this topic relevant to the product or a defect? (Confidence)
	Words	**Weight**	**Words**	**Weight**	**Words**	**Weight**	
1	Sleep	0.029	Pain	0.028	Bed	0.027	Yes (0.95)
	Pad	0.027	Night	0.021	Use	0.016	
	Topper	0.014	Back_Pain	0.014	Hip	0.013	
	Help	0.013					
2	Topper	0.035	Sleep	0.027	Bed	0.023	Yes (0.95)
	Night	0.017	Get	0.016	Cool	0.016	
	Hot	0.014	Purchase	0.013	Bought	0.012	
	memory_foam	0.012					
3	Bed	0.035	Topper	0.026	Smell	0.023	Yes (0.94)
	Hour	0.021	Like	0.018	inch	0.017	
	Size	0.016	Open	0.016	Air	0.016	
	Order	0.015					
4	Back	0.036	Return	0.027	Box	0.022	Yes (0.95)
	Topper	0.020	Would	0.019	One	0.019	
	Try	0.015	Review	0.011	Amazon	0.010	
	Disappoint	0.010					
5	Soft	0.036	Topper	0.026	Bed	0.022	Yes (0.95)
	Sink	0.020	Support	0.020	Like	0.019	
	Foam	0.018	Firm	0.014	Really	0.014	
	Body	0.014					

TABLE 5 | Summary of test cases.

Summary of test cases

Test cases	Relevance (Human Label)	Topic CV	Average confidence
Test case one	5 out of 5 topics are related to a furniture product	0.4930	0.948
Test case two	4 out of 5 topics are related to a furniture product	0.3583	0.950
Test case three	5 out of 5 topics are related to a furniture product	0.5162	0.948
Test case four	5 out of 5 topics are related to a furniture product	0.4256	0.950
Test case five	5 out of 5 topics are related to a furniture product	0.3959	0.944
Test case six	3 out of 5 topics are related to a furniture product	0.3503	0.946
Test case seven	4 out of 5 topics are related to a furniture product	0.3698	0.930
Test case eight	5 out of 5 topics are related to a furniture product	0.3942	0.950

TABLE 6 | Hypothesis test for LDA topic model.

One sample sign test for median

H_0: η = 0.7

H_1: η > 0.7

Sample (N)	8	Median	5
95% confidence interval for proportion	(3.936, 5)	*p*-value	0.035

the eight products. The following **Table 6** shows the result of the one-sample sign test for the median.

The test products have a mean of 4.375 out of five topics and a median of five out of five topics evaluated as relevant to a furniture product or problematic product information. They generated a mean topic CV of 0.4129 with all of them more than 0.3, which indicates the words in topics are somewhat coherent. Some test products generated a topic CV of higher than 0.5, which indicates that they have a good coherence between topics. This was tested with a one-sample sign test to evaluate if the model can retrieve 70% of the topic relevant to a furniture product or problematic product information. With the *p*-value at 0.035, the hypothesis test successfully rejected the null hypothesis, and thus, the LDA model can successfully retrieve key product defect topics.

CONCLUSION

This research explored the best method to provide insurers and manufacturers with an early view into product defects. While there is a large number of OCRs available on social media and

e-commerce sites, it is difficult for insurers and manufacturers to manually inspect those OCRs for defective information, which will delay product recall. This research has demonstrated a novel predictive model using RNN and LDA topic models to extract product defects from OCRs in a fast and automatic way. As the time to recall action increases during a product defect event, the recovery will be more challenging [17]. This new predictive model provides them with an early risk analysis on defective products and defect detection mechanisms.

In summary, this research has proposed and constructed two RNN classifiers and one LDA topic to determine if these models can retrieve product defect information. The first RNN model was able to identify negative OCRs, where most of these OCRs consist of complaints and issues about the products. The second RNN model was able to identify if the OCRs consist of one of the following information labels about the products' defect: "Manufacturing defect," "Problematic design or quality," and "Bad customer service." The Latent Dirichlet Allocation (LDA) model successfully combined the first and the second RNN models to retrieve key defect information on OCRs that were negative and mentioned product defect. This combined model was able to locate the keywords of the problems and issues that customers mentioned the most in their OCRs, which in turn achieved the goal of providing the insurers and manufacturers with an early view into potential product defect events.

DISCUSSION AND FUTURE WORK

The purpose of this research was to investigate the possibility of using probabilistic models for online customer reviews to retrieve product defect information and provide insurers and manufacturers with insight into the defects. Other than the two RNN models discussed earlier, this research also attempted to build a third RNN model to identify what kind of defect type ("Manufacturing defect," "Problematic design or quality," and "Bad customer service.") was mentioned in the OCRs. Due to the similarity of the use of words among the OCRs mentioning product defects, the RNN model was not able to classify OCRs down to the defect type. It was overfitted with the classification type that has the heaviest weight. The assumption is that the root cause is because the words used in all three classes of labels were so similar that the model was unable to separate the

probability space. The initial assumption was that the model could identify the difference among these words, but that was proven not to be the case. This limitation with this RNN approach provided an excellent space to use the LDA model. The LDA model, combined with the first and the second RNN models, successfully retrieved key information on OCRs that are negative and mention product defects.

Future research may consider improving and refining the accuracy of both the RNN classifiers and the LDA topic model and addressing the main outstanding problem identified in this research: the RNN classifier model for identifying defect types. RNN classifiers may be improved by adding multiple LSTM layers to construct a deeper neural network architecture. The accuracy may be improved this way, although the training time may increase. RNN classifiers may also be improved by training on a larger supervised dataset, while this would involve more manual labor work. The other neural network architectures such as Convolutional Neural Network layers can also be explored in future research for increasing the performance of identifying defective products. The deficiency for the RNN classifier that was attempted to identify defect types could be solved by constructing term frequency word lists for each defect type. Abrahams et al. have developed a set of smoke words, specifically for auto defect and auto safety issues, to identify whether OCRs mentioned defect or safety issues [24]. This method may be harder for particular classifiers to differentiate "Manufacturing defect" from "Problematic design or quality," since they have a very similar use of words in the OCRs.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

TF was responsible for running all the initial set of experiments and dataset preparation. SS and JF were the technical advisors for driving this project to completion.

REFERENCES

1. Ahsan K. Trend Analysis of Car Recalls: Evidence from the US Market. *Ijmvsc* (2013). 4(4):1–16. doi:10.5121/ijmvsc.2013.4401
2. Jeong E-Y. Samsung's Conclusions on Galaxy Note 7 Recall Backed up by Regulator. *Wall Street J* (2017). Retrieved from https://www.wsj.com/articles/samsungs-conclusions-on-galaxy-note-7-recall-backed-up-by-regulator-1486348636 (Accessed February, 2020).
3. Allianz Global Corporate & Specialty SE. AGCS Product Recall Report (2017). Retrieved from https://www.agcs.allianz.com/content/dam/onemarketing/agcs/agcs/reports/AGCS-Product-Recall-Report.pdf (Accessed March, 2020).
4. Mack R. San Francisco: CNET (2016). Another Samsung Galaxy Note 7 up in Smoke as Users Ignore Recalls. Retrieved from https://www.cnet.com/news/

another-samsung-galaxy-note-7-up-in-smoke-recalled-phone-not-put-down-data-shows/ (Accessed February, 2020).
5. Tibken S, and Cheng R. Samsung Answers Burning Note 7 Questions, Vows Better Batteries. San Francisco: CNET (2017). Retrieved from https://www.cnet.com/news/samsung-answers-burning-note-7-questions-vows-better-batteries/ (Accessed March, 2020).
6. Bleaney G, Kuzyk M, Man J, Mayanloo H, and Tizhoosh HR. Auto-detection of Safety Issues in Baby Products. *Lecture Notes Comp Sci Recent Trends Future Tech Appl Intelligence* (2018). 505–16. doi:10.1007/978-3-319-92058-0_49
7. Lee T, and Bradlow E. Automated Marketing Research Using Online Customer Reviews. *SSRN J* (2010) 48:881–894. doi:10.2139/ssrn.1726055
8. Suryadi D, and Kim HM. A Data-Driven Approach to Product Usage Context Identification from Online Customer Reviews. *J Mech Des* (2019). 141(12): 121104. doi:10.1115/1.4044523

9. CPSC. *U.S. Consumer Product Safety Commission*. Bethesda, MD: About CPSC (2019). Retrieved from https://www.cpsc.gov/About-CPSC (Accessed February, 2020).

10. Pruitt SW, and Peterson DR. Security price Reactions Around Product Recall Announcements. *J Financial Res* (1986). 9(2):113–22. doi:10.1111/j.1475-6803.1986.tb00441.x

11. Jarrell G, and Peltzman S. The Impact of Product Recalls on the Wealth of Sellers. *J Polit Economy* (1985). 93(3):512–36. doi:10.1086/261313

12. Dawar N, and Pillutla MM. Impact of Product-Harm Crises Onbrand Equity: The Moderating Role of Consumer Expectations'. *Marketing Res* (2000). 37(May):215–26. doi:10.1509/jmkr.37.2.215.18729

13. Matos CA, and Rossi CA. Consumer Reaction to Product Recalls: Factors Influencing Product Judgement and Behavioral Intentions. *Int J Consumer Stud* (2007). 31(1). doi:10.1111/j.1470-6431.2006.00499.x

14. Lyles MA, Flynn BB, and Frohlich MT. All Supply Chains Don't Flow through: Understanding Supply Chain Issues in Product Recalls. *Manag Organ Rev* (2008). 4(2):167–82. doi:10.1111/j.1740-8784.2008.00106.x

15. Beamish PW, and Bapuji H. Toy Recalls and China: Emotion vs. Evidence. *Manag Organ Rev* (2008). 4(2):197–209. doi:10.1111/j.1740-8784.2008.00105.x

16. Berman B. Planning for the Inevitable Product Recall. *Business Horizons* (1999). 42(2):69–78. doi:10.1016/s0007-6813(99)80011-1

17. Hora M, Bapuji H, and Roth AV. Safety hazard and Time to Recall: The Role of Recall Strategy, Product Defect Type, and Supply Chain Player in the U.S. Toy Industry. *J Operations Manag* (2011). 29(7-8):766–77. doi:10.1016/j.jom.2011.06.006

18. Siomkos GJ, and Kurzbard G. The Hidden Crisis in Product-harm Crisis Management. *Eur J Marketing* (1994). 28(2):30–41. doi:10.1108/03090569410055265

19. Souiden N, and Pons F. Product Recall Crisis Management: the Impact on Manufacturer's Image, Consumer Loyalty and purchase Intention. *Jnl Prod Brand Mgt* (2009). 18(2):106–14. doi:10.1108/10610420910949004

20. Karakaya F, and Ganim Barnes N. Impact of Online Reviews of Customer Care Experience on Brand or Company Selection. *J Consumer Marketing* (2010). 27(5):447–57. doi:10.1108/07363761011063349

21. Dellarocas C, Zhang X, and Awad NF. Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures. *J Interactive Marketing* (2007). 21(4):23–45. doi:10.1002/dir.20087

22. Barton B. Ratings, Reviews & ROI. *J Interactive Advertising* (2006). 7(1):5–50. doi:10.1080/15252019.2006.10722125

23. Chan NL, and Guillet BD. Investigation of Social Media Marketing: How Does the Hotel Industry in Hong Kong Perform in Marketing on Social Media Websites?. *J Trav Tourism Marketing* (2011). 28(4):345–68. doi:10.1080/10548408.2011.571571

24. Abrahams AS, Jiao J, Wang GA, and Fan W. Vehicle Defect Discovery from Social media. *Decis Support Syst* (2012). 54(1):87–97. doi:10.1016/j.dss.2012.04.005

25. Zhang X, Qiao Z, Ahuja A, Fan W, Fox EA, and Reddy CK. Discovering Product Defects and Solutions from Online User Generated Contents. In: The World Wide Web Conference on - WWW 19 Editors L. Liu and R. White (2019), San Francisco, CA, United States, 13 May, 2019–17 May, 2019, (New York: . ACM). doi:10.1145/3308558.3313732

26. El-Dehaibi N, Goodman ND, and Macdonald EF. Extracting Customer Perceptions of Product Sustainability from Online Reviews. *J. Mech. Des.* (2019). 141:121103. doi:10.1115/detc2019-98233

27. Rezaeinia SM, Rahmani R, Ghodsi A, and Veisi H. Sentiment Analysis Based on Improved Pre-trained Word Embeddings. *Expert Syst Appl* (2019). 117:139–47. doi:10.1016/j.eswa.2018.08.044

28. Mikolov T, Sutskever I, Chen K, Corrado G, and Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Red Hook, NY, USA: Curran Associates Inc. (2013). 2. p. 3111–9. doi:10.5555/2999792.2999959

29. Ren Y, Wang R, and Ji D. A Topic-Enhanced Word Embedding for Twitter Sentiment Classification. *Inf Sci* (2016). 369:188–98. doi:10.1016/j.ins.2016.06.040

30. Tang D, Wei F, Yang N, Zhou M, Liu T, and Qin B. Learning Sentiment-specific Word Embedding for Twitter Sentiment Classification. *Proc 52nd Annu Meet Assoc Comput Linguistics* (2014). 1:1014–1023. doi:10.3115/v1/p14-1146

31. Blei DM, Ng AY, and Jordan MI. Latent Dirichlet Allocation. *J Machine Learn Res* (2003). 3:993–1022. doi:10.1162/jmlr.2003.3.4-5.993

32. Pritchard J, Stephens M, and Donnelly P. *Inference of Population Structure Using Multilocus Genotype Data*. Rockville, MD: Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature (2003). doi:10.3410/f.1015548.197423

33. Santosh DT, Babu KS, Prasad S, and Vivekananda A. Opinion Mining of Online Product Reviews from Traditional LDA Topic Clusters Using Feature Ontology Tree and Sentiwordnet. *Int J Edu Manag Eng* (2016). 6(6):34–44. doi:10.5815/ijeme.2016.06.04

34. Zhai CX, and Massung S. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. New York, NY: ACM Books (2016). doi:10.1145/2915031

35. Titov I, and Ryan M. *A Joint Model of Text and Aspect Ratings for Sentiment Summarization*. ACL/HLT (2008).

36. Calheiros AC, Moro S, and Rita P. Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling. *J Hospitality Marketing Manag* (2017). 26(7):675–93. doi:10.1080/19368623.2017.1310075

37. Han J, Kamber M, and Pei J. *Data Mining: Concepts and Techniques*. Waltham, MA: Morgan Kaufmann (2012). p. 398.

38. Liu P, Qiu X, and Huang X. Recurrent Neural Network for Text Classification with Multi-Task Learning. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence IJCAI'16 Editors S. Kambhampati, New York, United States, 9–15 July 2016. (Palo Alto, CA, United States: . AAAI Press) (2016). p. 2873–9. doi:10.5555/3060832.3061023

39. Rumelhart DE, Hinton GE, and Williams RJ. Learning Representations by Back-Propagating Errors. *Nature* (1986). 323(6088):533–6. doi:10.1038/323533a0

40. Pearlmutter. Learning State Space Trajectories in Recurrent Neural Networks. In: International Joint Conference on Neural Networks Editors S.-I. Amari, Washington, DC, Unite States, June 18–22, 1989, (Washington, DC: IEEE) (1989). doi:10.1109/ijcnn.1989.118724

41. Bengio Y, Simard P, and Frasconi P. Learning Long-Term Dependencies with Gradient Descent Is Difficult. *IEEE Trans Neural Netw* (1994). 5(2):157–66. doi:10.1109/72.279181

42. Hochreiter S. The Vanishing Gradient Problem during Learning Recurrent Neural Nets and Problem Solutions. *Int J Unc Fuzz Knowl Based Syst* (1998). 06(02):107–16. doi:10.1142/s0218488598000094

43. Graves A. Offline Arabic Handwriting Recognition with Multidimensional Recurrent Neural Networks. *Guide to OCR for Arabic Scripts* (2012). 297–313. doi:10.1007/978-1-4471-4072-6_12

44. Tang J, Li H, Cao Y, and Tang Z. Email Data Cleaning. In: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining - KDD 05 Editors R. Grossman, Chicago, IL, United States, 21 August, 2005–24 August, 2005, (New York: Springer) (2005). doi:10.1145/1081870.1081926

45. Li J, Luong T, Jurafsky D, and Hovy E. When Are Tree Structures Necessary for Deep Learning of Representations?. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (2015). doi:10.18653/v1/d15-1278

46. Agarwal S, Godbole S, Punjani D, and Roy S. How Much Noise Is Too Much: A Study in Automatic Text Classification. In: Seventh IEEE International Conference on Data Mining Editors N. Ramakrishnan, O. R. Zaïane, Y. Shi, C. W. Clifton, and X. Wu, Omaha, NE, United States, 28–31 Oct 2007. (Washington, DC: . ICDM 2007) (2007). doi:10.1109/icdm.2007.21

47. Russell SJ, and Norvig P. *Artificial Intelligence: A Modern Approach*. Hoboken: Pearson (2011).

48. Chang J, Boyd-Graber J, Wang J, Gerrish S, and Blei DM. Reading tea Leaves: How Humans Interpret Topic Models. Neural Information Processing Systems (2009). Available from: https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf (Accessed May, 2020). doi:10.1145/1557019.1557044