



Marketing Science Institute Working Paper Series 2024

Report No. 24-102

Screening Consumer Complaints for Recall Management: A Topic Model for Decision Automation

Wen Shi, Yujie Qu and Jia Liu

“Screening Consumer Complaints for Recall Management: A Topic Model for Decision Automation” © 2024

Wen Shi, Yujie Qu and Jia Liu

MSI Working Papers are Distributed for the benefit of MSI corporate and academic members and the general public. Reports are not to be reproduced or published in any form or by any means, electronic or mechanical, without written permission.

Screening Consumer Complaints for Recall Management: A Topic Model for Decision Automation

Wen Shi Yujie Qu Jia Liu*

November 7, 2023

Abstract

While consumer complaints are recognized as the primary catalyst for product recalls in numerous sectors with high recall rates (such as automobiles, food, beverages, and pharmaceuticals), both firms and regulatory bodies face challenges due to limited human and technological resources when it comes to screening these complaints for trend analysis. Addressing this gap, we introduce a semi-parametric topic model named the hierarchically dual Pitman-Yor process (HDPYP). The HDPYP is designed to automatically process and analyze vast volumes of consumer complaints alongside their associated recall statements. The HDPYP not only extracts defect-related topics but also predicts the significance of each consumer complaint and forecasts the topic distribution of subsequent recall statements. We apply the HDPYP using consumer complaint datasets and vehicle recall data from the U.S. automobile sector. Our findings demonstrate the value of the HDPYP to aid firms and regulators in crucial decision-making processes, such as pinpointing pivotal consumer complaints warranting further examination (or those deemed “summary-worthy” in subsequent recall statements), identifying product defects, and forecasting recall occurrences in advance. Furthermore, by integrating the outputs of the HDPYP with Large Language Models (LLMs), regulators can efficiently and effectively review and authenticate the recall statements submitted by firms.

Keywords: product recalls, consumer complaints, topic modeling, prediction analysis, large language models, big data analytics

*Wen Shi is a Professor of Management Science and Information Management, School of Business, Central South University, China, (email: wenshi@csu.edu.cn). Yujie Qu is a Ph.D. Candidate in Management Science and Information Management, School of Business, Central South University, China, (email: jsqyj3332@csu.edu.cn). Jia Liu is an Associate Professor of Marketing, HKUST, HK (email: jiali@ust.hk). This paper is part of the second author’s doctoral dissertation. All the authors contribute equally. The authors would like to thank the seminar participants at CUHK, Cornell Young Scholar Camp, the 2022 Marketing Science Conference for their feedback.

1 Introduction

Consumer reviews and complaints provide insights into buyers' authentic and representative opinions, serving as a pivotal strategic resource for firms to improving customer experiences and making business decisions (Mejia et al., 2021). Specifically, consumer complaints offer a cost-effective way for firms to identify potential issues and prepare for the potential crises arising from product recalls. The current Good Manufacturing Practice (GMP), as mandated by the World Health Organization, emphasizes the necessity for firms to thoroughly comprehend relevant concerns in order to optimize their complaint handling and recall systems (ECA Academy, 2023). In recall-intensive sectors, such as automobiles, food, beverages, and pharmaceuticals, almost all recalls, whether voluntary or mandatory, stem from consumer complaints directed at the firms (John O'Dell, 2013). Swift action on product recalls is crucial; any delay can escalate risks and amplify financial damages. For example, the "don't tell" culture at the General Motors (GM) during the significant ignition switch crisis was linked to 124 deaths and 275 injuries, culminating in a \$900 million fine (Lareau, 2019).

To enhance recall management and reduce safety risks, regulators actively encourage consumers to report potential defects. For instance, the U.S. Food and Drug Administration (FDA) established MedWatch to gather data on adverse events in healthcare. Similarly, the U.S. National Highway Traffic Safety Administration (NHTSA) introduced the SaferCar App, aiming to disseminate safety information to the general public. These collected consumer complaints provide regulators with insights into a company's quality assurance practices, allowing them to determine any potential legal violations (Raval, 2020). However, the FDA has faced criticism for its perceived slow reactions to complaints, particularly those concerning contaminated formula and other vital health, safety, and nutritional matters. The NHTSA, too, has been criticized for its apparent failure to thoroughly analyze data and pinpoint threats (Murray, 2022). A primary factor contributing to these inefficiencies is the daunting task of filtering crucial complaints from a vast influx. For instance, the NHTSA had fewer than 10 individuals responsible for manually screening 77,000 safety complaints (GAO, 2016), relying heavily on basic data statistics (ConsumerReports.org, 2015). Consequently, these initial screeners are burdened with an overwhelming workload, and some complaints might have been given mere seconds of consideration (NHTSA, 2015, page 15). China's regulator for product quality and safety, the State Administration for Market Regulation (SAMR), faces similar challenges. A modest team of 3-4 technicians have to manually review more than 40,000 vehicle-related complaints annually, potentially leading to recall delays.¹

¹The authors carried out several semi-structured interviews with SMAR officials. The ensuing statistics and details were provided by the pertinent department within SMAR.

Regulators urgently need more advanced statistical approaches to effectively comprehend and address consumer complaints (NHTSA, 2015, page 13). In 2021, the NHTSA allocated a budget of \$37 million for such enforcement, and \$28 million was earmarked for trend analyses using consumer complaint datasets (DOT, 2020). There is also a burgeoning trend towards leveraging artificial intelligence (AI) to transform the complaint management process across industries (Foster et al., 2021). McKinsey & Company posits that the capacity to promptly foresee and execute timely product recalls can curtail warranty expenses. This is achieved by helping businesses preempt potential issues through knowledge transfer, risk evaluation, and rigorous testing and validation (Aragon et al., 2019). To achieve this, companies can deploy automated technologies to continuously monitor and assess consumer complaints. This allows them to discern when in-depth investigations or response management are necessary and when there is a need to preemptively notify their quality assurance units of looming issues. Deploying such technological solutions is estimated to potentially save the automotive industry a staggering \$50 billion annually (Phukan, 2017). Despite the pressing demand for such advanced tools, academic literature predominantly focuses on institutional factors leading to recalls (e.g., Liu et al., 2016; Thirumalai and Sinha, 2011; Singh and Grewal, 2023) and the consequences of these recalls (e.g., Chen et al., 2009; Che et al., 2023; Liu and Shankar, 2015).

This paper aims to fill in this significant research gap by developing a semi-parametric topic model, named the hierarchically dual Pitman-Yor process (HDPYP), that can proficiently and effectively pinpoint essential consumer complaints without relying on labeled datasets. The HDPYP can autonomously handle vast structured (e.g., metadata such as fatalities, injuries, and product features) and unstructured data (e.g., text) from consumer complaints and affiliated recall statements. A trained HDPYP can discern defect themes/components inherent to each consumer complaint, and predict the relative significance of each complaint concerning its eventual recall statement (if recalled). Our methodological approach draws inspiration from two primary observations. First, there exists a severe lack of labeled consumer complaints datasets suitable for training machine learning models dedicated to the screening of such complaints. Specifically, neither manufacturers nor regulators have meticulously annotated which complaints or to what extent each complaint influenced a subsequent product recall. This largely explains the limited methodological research in this area, despite its enduring significance. Second, a product recall is often precipitated by a handful of complaints amidst a sea of many (Clifford Atiyef, 2014; Coolidge, 2018). This phenomenon also aligns with the renowned Pareto principle in management (Schmittlein et al., 1993). To ensure efficient and scalable model inference, we develop a blocked and collapsed Gibbs sampler and also introduce a novel method to determine optimal model hyper-parameters.

We apply the HDPYP in consumer complaints and vehicle recalls sourced from the NHTSA. This sector is particularly pertinent as it surpasses all others in the frequency of recalls. Over the past decade, upwards of 280 million vehicles have been recalled, resulting in an estimated annual economic setback exceeding \$230 billion (NHTSA, 2019). We demonstrate the efficacy of the HDPYP in gleaning valuable insights, spanning defect topics, document topic distribution, determinants of complaint significance, and the importance of each complaint. Moreover, our results illuminate the potential of the HDPYP in facilitating two crucial decision-making facets in recall management: the prediction of recall occurrences (inclusive of defect components) and the composition/verification of recall statements. The proactive prediction of recall incidents is vital for mitigating the adverse ramifications of product defects (Aragon et al., 2019). Our analyses reveal that, compared to conventional methods that don't differentiate between complaints or other heuristic techniques, utilizing the relative complaint importance scores derived from the HDPYP enhances the prediction accuracy of recall occurrences by approximately 14% ($p < 0.05$), and improves the precision of identifying the specific recalled components by roughly 23% ($p < 0.05$). To further illustrate the tangible benefits of enhanced recall prediction accuracy, we delve into a case study centered on a selection of notable and economically consequential recall events within the automotive sector. Our results indicate that, compared to benchmark models, the HDPYP is capable of anticipating recall incidents at least a full year earlier, which could have decreased business losses and protected customers from further property damage or loss of life.

When a recall is enacted, the implicated manufacturer must issue a public statement detailing the defective products and the proposed remedy, serving as a vital communication tool in crisis scenarios. Drafting an apt recall statement demands significant time and effort from manufacturers as they meticulously review and aggregate all associated complaints (NHTSA, 2019). Once crafted, this statement is sent to the regulator, who then cross-references its contents with related consumer complaints. Given the sheer volume of data, this verification process can often be inefficient and ineffective, according to regulators at SMAR in China. In our empirical analysis, we demonstrate how the HDPYP can streamline this arduous process. Firstly, the HDPYP can predict which defect topics or consumer complaints should be prioritized or deemed "summary-worthy" by those drafting the recall statement. Secondly, the HDPYP can be integrated with Large Language Models (LLMs), enabling the automatic crafting of recall statements. We exemplify this utility using the state-of-the-art LLM, the Text-To-Text Transfer Transformer (T5) (Raffel et al., 2020). Our findings indicate that feeding T5 with the most pertinent consumer complaints, as determined by the HDPYP, yields superior recall statements compared to other benchmark methods (such as using all consumer complaints, or solely relying on complaints associated

with fatalities or injuries). In sum, leveraging the complaint importance weights predicted by the HD-PYP results in significantly enhanced outcomes in terms of defect identification, recall prediction, and statement drafting, as opposed to using rudimentary heuristics or not differentiating complaints at all.

The rest of this paper is organized as follows. We first review the relevant literature in Section 2. Our proposed topic model, along with the inference algorithm and the optimization of model hyperparameters, are introduced in Section 3. We describe our empirical study in Section 4, followed by the model substantive outputs in Section 5. Next, we showcase the applications of predicting recall incidence and of drafting recall statements in Section 6. We conclude with a discussion in Section 7.

2 Relevant Literature

Our research lies at the intersection of marketing, operation management, and machine learning. We provide an overview of relevant studies on product recalls and on analyzing user-generated content (UGC) using topic modeling, and highlight our contributions to these domains.

Recall procedures are notoriously protracted, so there is a considerable interest in identifying the crucial triggers of recalls. The current empirical studies mostly center on institutional factors: the geographical configuration of the supply chain (Hora et al., 2011), specific attributes like the origin of reports or the entities involved (Ni and Huang, 2018), and the gender of the directors (Wowak et al., 2021). Some analytical models have also emerged to determine the optimal recall timing through dynamic programming, aiming to cut down on the recalling firm's projected expenses (e.g., Sezer and Haksöz, 2012; Mukherjee et al., 2022). In the computer science literature, Zhang et al. (2016) introduced an unsupervised probabilistic model adept at identifying product flaws using consumer complaints. However, its focus leans more toward descriptive rather than predictive. Collectively, the existing studies mostly illuminate the supply side, namely manufacturers and suppliers, while little has been directed towards probing the consumer side, thus empowering companies and regulators to act preemptively. To the best of our knowledge, our paper is pioneering in its consumer-centric, data-driven approach to scrutinize product recalls. Our advanced methodology not only offers an automated interpretation of consumer complaints but also paves the way for predicting recalls, both in terms of their likelihood and specifics.

There has been a growing number of studies in marketing on analyzing UGC (e.g., Archak et al., 2011; Lee and BradLow, 2011; Netzer et al., 2012; Liu et al., 2016; Berger et al., 2020). To extract meaningful information from UGC, some studies utilize probabilistic topic modeling methods to cluster items by content similarity or word co-occurrence. The most commonly used topic modeling method

is based on the latent Dirichlet allocation (LDA) (Blei et al., 2003). For example, LDA has been used to process online chatter (Tirunillai and Tellis, 2014), social tags (Nam et al., 2017), online reviews (Büschken and Allenby, 2016), online search (Liu and Toubia, 2018), and social media content (Zhong and Schweidel, 2020). Recent marketing studies adopt Poisson factorization to model UGC. For example, Toubia (2020) studied the generation of creative documents (e.g., abstracts and synopses); Liu et al. (2021) developed a content-based search model that establishes the links between the volume of user search query and click-through rate, to their content preference on search engines.

Note that most existing studies use topic models to examine unilateral UGC, while research on bilateral content posted by distinct agents remain rare. With the development of platform economy, there is a growing interest in analyzing bilateral reviews (Chen et al., 2021). To the best of our knowledge, the HDPYP is the first to explicitly capture the relationships of one type of documents (e.g., consumer complaints) on the generation of a different type of documents (e.g., recall statements by firms) that are created by distinct parties. Among the existing topic models, the LDA-based topic model proposed in Liu and Toubia (2018) is the closest to our proposed HDPYP in terms of the high-level model structure. Liu and Toubia (2018) enables semantic relationship between search queries and search results on search engine result page in which multiple search queries could be linked with the same search result and the same search query could be linked with multiple search results; the topic distributions of different documents are aggregated using arithmetic mean. In comparison, our proposed HDPYP allows explicit mapping from multiple documents to one corresponding summary document, and also quantifies the importance of each document to the summary document with interpretable insights. These are nontrivial extensions that require a significantly different modeling approach.

We achieve these objectives by building our topic model based on the Pitman-Yor process (PYP), a well-known nonparametric Bayesian modeling technique (which is also known as a two-parameter Poisson-Dirichlet process) for complex topic modeling (Lim et al., 2016). PYP not only captures the possible topic relationships of distinct documents in a corpus, but also fits the data better than other stochastic processes due to its power-law property (Fan et al., 2017). This is particularly important for many real-world applications because most words/items appear rarely in the corpus. PYP has already been a popular topic modeling technique in several fields, including genetics (Ni et al., 2018), epidemiological tracking (Caron et al., 2017), and online news topic detection (Fan et al., 2021). Our paper is the first to apply PYP in management decision problems.

Our proposed HDPYP has a significantly different structure, compared with the existing extensions of the basic PYP, mainly the hierarchical PYP (HPYP) (Lim et al., 2016) and the compound PYP (CPYP)

(Wood and Teh, 2009). HPYP focuses merely on the internal structure of one type of documents, which is formed by many segments. In contrast, our proposed HDPYP captures the semantic mapping between two types of documents (i.e., multiple complaints correspond to one recall statement), as well as the internal structure of each type of documents. The CPYP extends the single base measure by including multiple input distributions with their weights which are sampled from a beta distribution. In contrast, our proposed HDPYP not only takes the effect of metadata into the weights, but also extends the weights to be multidimensional. Therefore, the HDPYP extends the existing hierarchical topic models in several directions to capture the more complex characteristics of our study context than the existing methods.

3 HDPYP

3.1 Model Specifications

The HDPYP is a semiparametric topic model for bag-of-words data in which the topic distributions of many documents (e.g., consumer complaints) have different effects on the topic distribution of the resulting *summary* document (e.g., a recall statement). We now describe the HDPYP using our empirical context, but it can be generalized to any corpus with a similar structure. Let V denote the total number of distinct words observed in recall statements and consumer complaints. The words in the vocabulary are indexed by $v \in \{1, 2, \dots, V\}$. Suppose that there is a collection of I recall statements. We let L_i denote the number of words in recall statement i , and $w_{i,l}^r$ denote the l th word token in recall statement i . Let J_i denote the number of complaints associated with recall i . For the j th complaint of recall i , we let $H_{i,j}$ denote its number of words and $w_{i,j,h}^c$ denote its h th word token. Figure 1 provides a graphical representation of the HDPYP.

Topics. Suppose that there are K different topics. We assume that consumer complaints and recall statements share the same set of topics, but they exhibit these topics with different intensities. Following the standard specification in the latent Dirichlet allocation (LDA) (Blei et al., 2003), we model each topic k as a distribution over all of the words in the vocabulary, $\phi_k \sim \text{Dirichlet}_V(\gamma)$, where γ is a V -dimensional vector that controls the sparsity of the word distribution.

Consumer Complaints. Conceptually, complaints associated with the same recall should include similar semantic information. Thus, we let μ_i denote the *parent* document-topic distribution of the complaints associated with recall i . We assume that μ_i follows a Dirichlet distribution, i.e., $\mu_i \sim \text{Dirichlet}_K(\alpha)$, where the hyper-parameter α is an K -dimensional vector. This is equivalent to the *base distribution* rooted in the one-parameter Dirichlet process (DP), which is a stochastic process that can be regarded as

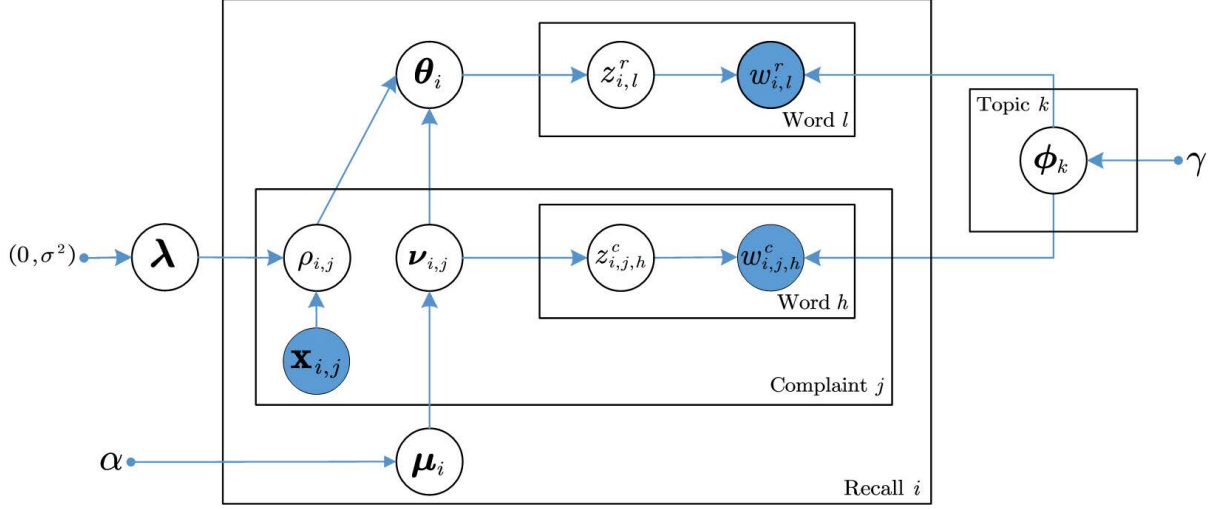


Figure 1: Graphical Representation of the HDPYP

an infinite dimensional generalization of the Dirichlet distribution. Next, we model the topic distribution $\mathbf{v}_{i,j}$ of each complaint j associated with recall i , on the basis of its parent distribution μ_i , using the PYP (Buntine and Hutter, 2012),

$$\mathbf{v}_{i,j} \sim \text{PYP}(a^v, b^v, \mu_i), \quad j = 1, 2, \dots, J_i, \quad (1)$$

where $a^v \in [0, 1)$ is a discount parameter and $b^v \in (-a, \infty)$ is a concentration parameter used in the DP. In the above specification, $\mathbf{v}_{i,j}$ is drawn from an infinite discrete probability distribution, consisting of an infinite set of atoms drawn from μ_i , with weights drawn from a two-parameter Poisson-Dirichlet distribution. The h th word in the j th complaint of recall i is generated similarly as in the LDA. We sample its topic indicator $z_{i,j,h}^c$, and then sample the observed word $w_{i,j,h}^c$ as follows:

$$z_{i,j,h}^c \sim \text{Categorical}_K(\mathbf{v}_{i,j}), \quad \text{and} \quad w_{i,j,h}^c \sim \text{Categorical}_V(\phi_{z_{i,j,h}^c}). \quad (2)$$

Recall Statements. We sample the topic distribution θ_i of recall i according to the topic distributions of its associated consumer complaints, i.e., $\{\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \dots, \mathbf{v}_{i,J_i}\}$. We let $\rho_i \triangleq (\rho_{i,1}, \rho_{i,2}, \dots, \rho_{i,J_i})$ denote the relative importance of each observed consumer complaint to the topic distribution of recall statement i . Next, we generate θ_i using the PYP as

$$\theta_i \sim \text{PYP}(a^{\theta_i}, b^{\theta_i}, \sum_{j=1}^{J_i} \rho_{i,j} \mathbf{v}_{i,j}), \quad (3)$$

where the weighted topic distributions of all of the associated complaints form the base distribution of the stochastic process. This specification leads to a hierarchical PYP with three layers for each recall

i , namely, the root layer μ_i , the middle layer $\{v_{i,1}, v_{i,2}, \dots, v_{i,J_i}\}$, and the high layer θ_i . Our model specification extends the hierarchical PYP proposed in Lim et al. (2016) by enabling the middle layer to have many different PYP nodes across observations. Finally, for the l th word in the i th recall, we sample its topic indicator $z_{i,l}^r$, and then sample the word $w_{i,l}^r$ as follows:

$$z_{i,l}^r \sim \text{Categorical}_K(\theta_i), \text{ and } w_{i,l}^r \sim \text{Categorical}_V(\phi_{z_{i,l}^r}). \quad (4)$$

To make an inference about the importance weights of future consumer complaints, we model ρ_i as a function of observed complaint-specific structured information (such as the automaker, fires, injuries, crashes, and deaths), denoted as a F -dimensional vector $x_{i,j}$ for complaint j associated with recall i . We specify ρ_i using a Dirichlet-multinomial regression model,

$$\rho_i \sim \text{Dirichlet}_{J_i} \left(\exp(\lambda^\top x_{i,1}), \exp(\lambda^\top x_{i,2}), \dots, \exp(\lambda^\top x_{i,J_i}) \right), \quad (5)$$

where the F -dimensional parameters λ capture the effects of these covariates on the importance of complaint j to recall statement i , λ are sampled from a Normal distribution with mean 0 and variance σ^2 . Note that although in prior ρ is only a function of the observed metadata in consumer complaints, the posterior distribution of ρ incorporates the textual information in all of the documents (see Web Appendix B). In this way, the trained Equation (5) leverages the semantic relationship of the textual information between consumer complaints and recall statements, thereby making better predictions of the importance weights of future consumer complaints instead of simply relying on the observed metadata.

3.2 Model Inference

Given the observed covariates and the contents of recall statements and consumer complaints, we need to estimate the following model parameters: $\{\lambda, \mu, v, \theta, \Phi, \rho, z^r, z^c\}$. We adopt the Markov chain Monte Carlo (MCMC) to obtain approximate inference. To improve efficiency, we cast the PYP nodes into the Chinese Restaurant process (CRP) metaphor (Blei et al., 2010; Lim et al., 2016), treating each PYP node as a restaurant and each word as a customer. To strike a good balance between computational storage and algorithm efficiency, we adopt the notion of a *table indicator* (Chen et al., 2011), which acts as an auxiliary variable and only requires access to the information about whether a customer opens a new table. The new representation of the HDPYP is described in Web Appendix A.

With table indicators, we employ a combination of the blocked and collapsed Gibbs samplers for

efficient model inference as in [Lim et al. \(2016\)](#). By integrating out the other parameters, our algorithm will iterate over table indicators and word-topic assignments from their posterior distributions. This sampler is blocked because these two types of parameters are updated simultaneously in each iteration. These integrated parameters can be approximated from the associated customer counts and table counts, using the expected values of their posterior distribution via sampling. Over the MCMC iterations, the algorithm only stores the information about customer counts and table counts associated with each restaurant, and therefore requires much less computational space, but at no loss of statistical inference efficiency. The details of our inference algorithm are provided in [Web Appendix B](#).

3.3 Optimization of Hyperparameters

The HDPYP has three types of hyperparameters: the number of topics, the discount and concentration parameters in the PYP (i.e., a and b), and the hyper-priors in the Dirichlet distributions (i.e., σ^2 , α , and γ).² Specifying appropriate model hyper-parameters is vital to topic models, as it may influence model fitting and robustness ([Wallach et al., 2009](#)). A common solution is to optimize one hyper-parameter at each time while keeping the others unchanged at a constant value (e.g., [Lim et al., 2016](#); [Liu and Toubia, 2018](#)). However, this approach is inefficient when the number of hyper-parameters is large, which is true in the HDPYP.

We propose a general and efficient solution to search for the optimal combination of model hyper-parameters. The basic idea is that we treat a topic model (e.g., the HDPYP) as a simulation model, with the inputs being the hyper-parameters and the output being the fitting indicator of interest (e.g., the perplexity score); we can then observe an output after the simulation takes a single run. We adopt the most popular metamodeling method – the stochastic Kriging (SK) method (also known as the Gaussian process) ([Ankenman et al., 2010](#)). Metamodels are an mathematical approximation of the input/output (I/O) relationship implied by the underlying simulation model ([Kleijnen, 2015](#)), and can be used to make prediction on unexplored parameter space efficiently. All of the technique details are provided in [Web Appendix C](#). We find that the SK metamodeling method only takes seconds to yield an estimate of the perplexity of a given parameter combination. To the best of our knowledge, our paper is the first to develop an efficient solution for optimizing hyper-parameters in the literature of topic modeling.

²For simplicity, we follow the standard approach in the topic modeling literature, and treat α and γ as a vector with the same scalar (i.e., the symmetric prior).

3.4 Applications

We illustrate how to apply the HDPYP in practice for various objectives using a concise logical flow shown in Figure 2. First, one has to obtain a collection of recalled products in a certain product category, and their associated recall statements and consumer complaints (along with some metadata). Second, the HDPYP is calibrated by this historical data (detailed in Section 3.2 and 3.3). Third, the estimated model parameters of the HDPYP can provide substantive insights, such as defect topics, topic distributions within a document and across documents, features of important complaints, etc. Fourth, for each newly received consumer complaint for any product model, one can apply the trained HDPYP to estimate its distribution of defect topics (using calibrated topics and calibrated Equation (1)), and update the importance weight of each consumer complaint associated with the same product model (using calibrated Equation (5)). These predicted complaint importance weights can be further leveraged for other managerial decisions, such as screening consumer complaints, summarizing potential defects, predicting recall incidence and defect components, and drafting/verifying recall statements. We will illustrate these applications with more details in our empirical study.

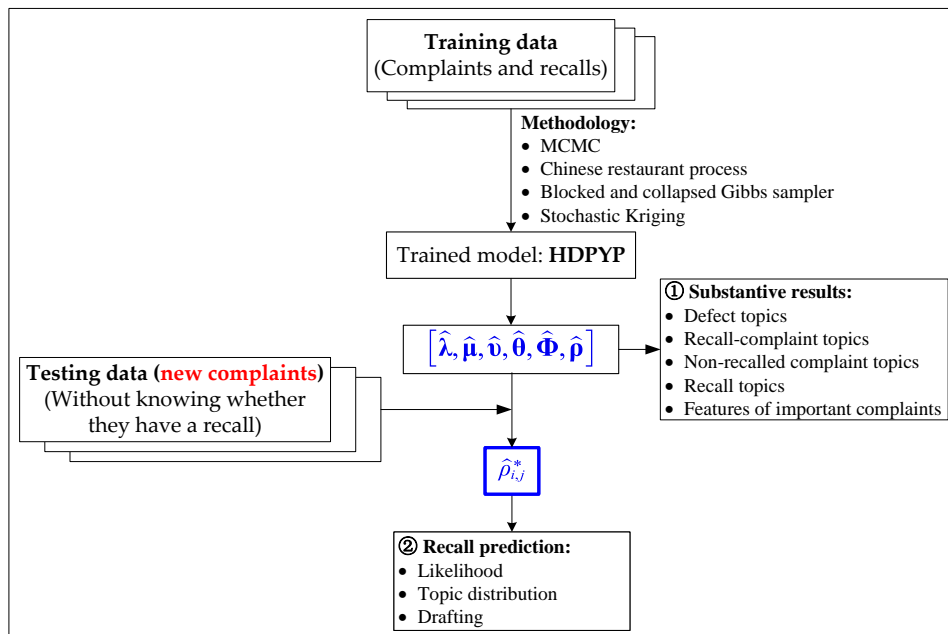


Figure 2: A Synopsis of Implementing HDPYP for Practitioners

4 Empirical Study

4.1 Research Context and Data

Our data sets, which have been a major data source for recall-related research, are obtained from a public U.S. platform launched by the NHTSA. As mentioned in the introduction, the NHTSA is responsible for overseeing vehicle safety in the U.S. by initiating recalls, monitoring the effectiveness of ongoing recalls, and maintaining various communication channels for consumers to submit complaints (e.g., phone, email, and website). When an automaker decides to withdraw a defective product through a recall, it must inform the NHTSA and provide a recall statement (containing a clear explanation of the recall issue, the safety consequences, and the risks associated with potential malfunctions) within a few (usually five) business days (Astvansh et al., 2022). Once a recall is warranted, the associated manufacturer must provide notice to owners of the affected vehicles or equipment within 60 days.


We acquired the recall reports in January 2020, including every recall between 2000 and 2019. The NHTSA has listed 26 defect components for vehicle recalls. The unit level of the automobile recalls is make-model-year (MMY). For each recall, the NHTSA provides a statement that describes the potential safety hazard, component(s) for the recall, along with MMY information (e.g., the vehicle maker, model, model year, beginning and ending manufacturing dates, number of units affected, and recall initiator). See an example of recall statement in Figure 3. In addition, we collected information associated with each MMY (i.e., body, cylinder, engine, drive type, and transmission) from *Teotalida.com*, the largest data provider in the automobile industry.

We also acquired safety-related complaint information from the NHTSA consumer-complaint database over the same time period.³ Each complaint report includes a textual description of the potential defect, along with structured metadata, including the complaint date, manufacturer, model, model year, possible defect component(s), whether a crash and/or a fire occurred, the number of injuries, the number of deaths, whether an anti-brake system was involved, the state where the consumer resides, and the source of the complaints (online or hotline⁴). The NHTSA has listed 29 components for consumers to choose from, along with an “other” option for non-specific components, and consumers can select multiple components. But this function is not common across similar platforms. See a sample consumer complaint in the second panel of Figure 3.

³Each valid complaint has a unique VIN code, so it is unlikely for a large number of false complaints to be filed by fundamentally different consumers who do not know each other.

⁴If a complaint is filed through the hotline, the NHTSA database also provides complete textual description.

2011
FORD FIESTA
4 DR FWD



June 13, 2012 NHTSA ID NUMBER: 10461636

Components: AIR BAGS, STRUCTURE

NHTSA ID Number: 10461636

Incident Date May 24, 2012

Consumer Location HOLLAND, MI

Vehicle Identification Number 3FADP4FJ9BM****

Summary of Complaint

CRASH	Yes	VEHICLE WAS IN A MINOR 'FENDER BENDER' DURING RUSH HOUR - STOP AND GO - TRAFFIC. AIRBAGS DID NOT DEPLOY AND DAMAGE TO THE 2011 FIESTA SES WAS FIXABLE. VERY MINIMAL DAMAGE (DENT) TO CAR IN FRONT OF MY VEHICLE. MY VEHICLE WENT FROM FIXABLE TO IGNITING AND A TOTAL LOSS. *TR
FIRE	Yes	
INJURIES	0	
DEATHS	0	


1 Affected Product ▾

Vehicle

MAKE	MODEL	YEAR
FORD	FIESTA	2011

(a) Complaint Report

2011
FORD FIESTA
4 DR FWD



October 5, 2012 NHTSA CAMPAIGN NUMBER: 12V488000

Passenger Side Curtain Air Bag/FMVSS 208

An occupant in the right rear seating position will not have coverage from the side curtain air bag in a side impact collision when the front passenger seat is empty, increasing the risk of injury to the right rear occupant.

NHTSA Campaign Number: 12V488000

Manufacturer Ford Motor Company

Components AIR BAGS

Potential Number of Units Affected 154,471

Summary

Ford is recalling certain model year 2011-2013 Fiesta vehicles, manufactured from November 3, 2009 through September 21, 2012. The vehicles fail to comply with Federal Motor Vehicle Safety Standard (FMVSS) No. 208, "Occupant Crash Protection." The passenger side curtain air bag will not deploy in the event of a side impact collision when the front passenger seat is empty. Although the side curtain air bag system was designed to suppress the side curtain air bag under this scenario, that information is not explained in the owner's guide for these vehicles as required by FMVSS No. 208.

Remedy

Ford will notify owners and dealers will reprogram the vehicle's software free of charge so that it no longer suppresses the passenger side curtain air bag when the front passenger seat is empty, and consistent with the description in the owner's guide. The safety recall began on October 26, 2012. Owner's may contact Ford at 1-866-436-7332.

Notes

Ford's recall campaign number is 12C26. Owners may also contact the National Highway Traffic Safety Administration Vehicle Safety Hotline at 1-888-327-4236 (TTY 1-800-424-9153), or go to www.safercar.gov.

3 Affected Products ▾

Vehicle

MAKE	MODEL	YEAR
FORD	FIESTA	2011-2013

(b) Recall Statement

Figure 3: An Example of Complaint Report and Recall Statement

4.2 Sample Selection

Following the prior literature (Astvansh et al., 2022; Singh and Grewal, 2023), we focus on the top 20 automakers (ranked by total sales during our observation period) for our empirical application. Even within the top-20 automakers, there are over 55,000 complaints annually, and each complaint requires meticulous analysis, rather than a simple skim reading. The top-20 automakers result in 3,660 unique MMYs. We further restrict our attention to the MMYs that have at least three associated complaints because too few complaints might not indicate a defect trend. If a MMY is recalled multiple times over our study period, we only include the observations up to the very first recall because multiple recalls of the same MMY may have different levels of correlation (Gao et al., 2022). Our final data set contains 62,353 complaints and 1,280 recalls, involving 1,693 unique MMYs.⁵ We conduct a standard procedure to pre-process all of the textual information in the consumer complaints and recall statements. We convert all of the words to lower case, remove all non-English characters and words with fewer than three characters, eliminate a standard list of English stop words, and apply word stemming.

4.3 Descriptive Statistics

Across these selected MMYs, about 24% were not recalled, indicating a low level of data imbalance. We report some summary statistics of these not-yet-recalled MMYs, these recalled MMYs, and their resulting recall statements in Table 1. Intuitively, we find that these non-recalled MMYs are associated with a larger number of complaints than the already-recalled MMYs because the former has a longer observation window. For the recalled MMYs, we find that on average, 45% of the words that appear in the recall statements also appear in their associated consumer complaints.

Table 1: Descriptive Statistics of the Corpus

	Complaints about Non-Recalled MMYs	Complaints about Recalled MMYs	Recall Statement about Recalled MMYs
Num. MMYs	413	1,280	1,280
Num. Unique Words	13,269	17,136	1,984
Num. Complaints per MMY	53.62 (95.05)	31.41 (53.45)	–
Num. Sentences per Document	7.14 (4.70)	7.09 (4.61)	4.37 (1.85)
Num. Words per Document	39.49 (28.57)	38.13 (28.46)	49.33 (26.91)

We now examine the proportion that each component is the defect component across recall statements and is the self-reported component across complaints. The distributions are shown in Figure

⁵For a small number of recall events, we find that multiple MMYs that have the same defect issue. In this case, we treat them as multiple recall statements, one for each MMY.

4. Note that the two sets of components are the same, except that the platform gives consumers four additional choices, which are shown in the last four bars in Figure 4: *other*, *traction control system*, *communications*,⁶ and *lane departure*. These components, however, rarely appear in consumer reports except *other*. One can see that most components have a small but significant appearance in recall statements and in consumer complaints. This observation suggests that a MMY could potentially be recalled or reported for many possible components, and there are no few components that can explain or predict most defects. In addition, there are significant differences in the component distributions across consumer complaints and across recall statements. While *air bags* is the top one component for manufacturers to issue vehicle recalls, consumers are much more likely to choose other components in filling complaints, including *power train*, *electrical system*, and *vehicle speed control*. Many other components also exhibit large differences, such as *fuel system*, *exterior lighting*, *equipment*, *seat belts*, and *engine*.

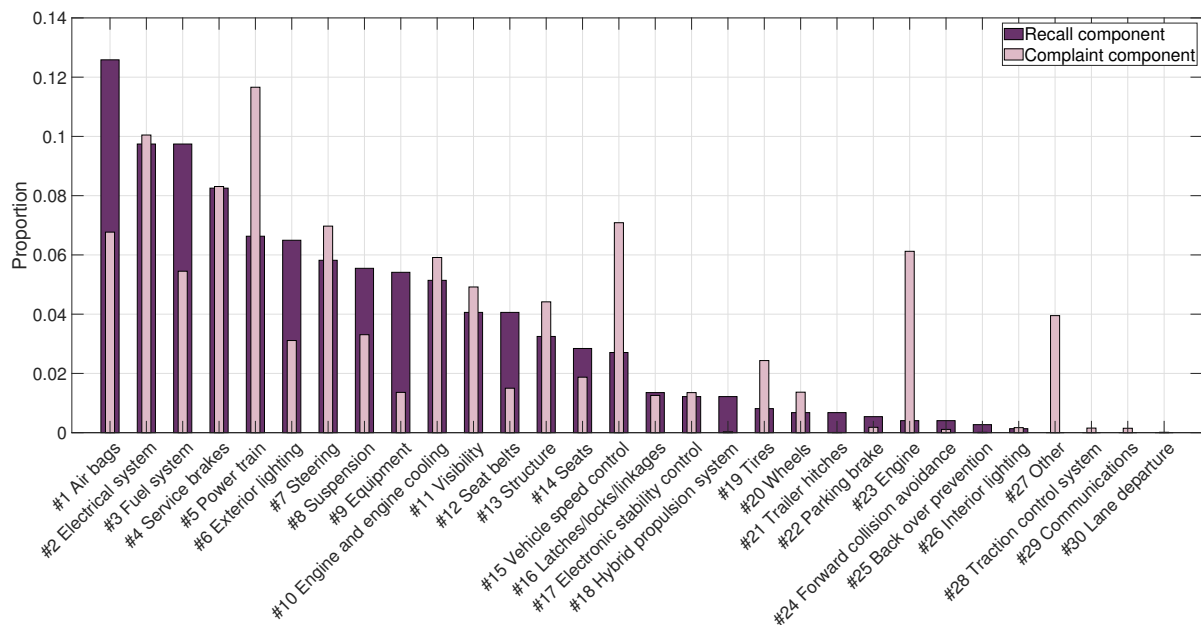


Figure 4: The Distribution of Defect Components in Complaints and Recall Statements

We formally evaluate this discrepancy at the level of recall statement. For each recalled MMY, we compute the proportion of times that each possible defect component was mentioned across all of its associated complaints, and then we identify the proportion for the components that matches with the actual defect components in the recall statements. We find that the average matching rate across MMYs is 0.106 with a standard deviation of 0.161. This low matching rate further confirms that consumers are generally not capable in providing categorical information on defect components. This is consistent

⁶Note that in complaint component *Communications* refers to either “Communications: Horn Assembly” or “Communications: Back Up Alarm.”

with the evidence from the U.S. government (see [NHTSA, 2015](#), page 5) that most of the complaints (between 50% and 70%) reported by consumers incorrectly identify the affected components.

Modern vehicles are technologically sophisticated, integrating numerous electronic, mechanical, and software components. The average consumer may lack the technical knowledge to accurately identify which specific component is causing an issue, especially when symptoms can overlap between components. Consumers might prioritize and report issues that have a more immediate and noticeable impact on their driving experience. Thus, a latent defect in an air bag might not be reported because it is not actively impacting their daily driving, even though it is a critical safety concern. Also, consumers are more likely to report symptoms they observe or are familiar with (like engine or power train), rather than the underlying defect or specific subsystems. Therefore, without expert knowledge, consumers may attribute problems to the most obvious or commonly known components.

Table 2: Comparing Complaints Associated with Recalled versus Not-Yet-Recalled MMYs

Variables	Coefficient of the recall	
	All MMYs	Recalled MMYs
Time gap between complaints	5.463*	4.281
Word count per complaint	0.597	-0.628
Positive sentiment per complaint	0.003	0.004
Negative sentiment per complaint	-0.006	0.024
Cum. num. crashes	0.092	0.035
Cum. num. fires	0.008	-0.031
Cum. num. injured	0.032	-0.019
Cum. num. death	0.004	0.009
Cum. num. top-10 state complaints	0.903*	0.113
Cum. num. anti-brake assembled	0.178	-0.324
Cum. num. online complaints	1.108*	0.497

Note. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Lastly, we examine whether the observed structured information of consumer complaints differs substantively between the recalled MMYs and the not-yet-recalled MMYs. Let Y_{it} denote a certain characteristic of MMY i at time t (at the level of year-month) based on its complaints received so far. Let $Recalled_{it}$ indicate whether the MMY i has already been recalled by time t . We consider a fixed-effect model of Y_{it} as

$$Y_{it} = \alpha + \beta Recalled_{it} + \xi_{f_{it}} + \gamma^T Z_i + \varepsilon_{it}, \quad (6)$$

where Z_i indicates all of the MMY-specific characteristics and let f_{it} indicates the lifetime of the MMY i by time t . The coefficient β is our key point of interest. We estimate the above regression model for 11 complaints related variables separately: the average time gap between complaints, the average word

count, the average sentiments based on the LIWC (Pennebaker et al., 2015), and seven cumulative count variables of the continuous information in the complaint metadata. We also use two sets of MMYs separately, i.e., all of the MMYs and only the recalled MMYs. The estimated β for all of the scenarios are shown in Table 2. We find that for most variables, there are either minimal or insignificant differences between the already-recalled and not-yet-recalled MMYs. This finding confirms the difficulty of identifying the MMYs that are at risk by merely relying on basic structured information in all of the consumer complaints. This further supports the importance of developing advanced NLP tools that can automatically process consumer complaints and estimate their importance to potential recalls.

4.4 Empirical Approach

When estimating the HDPYP using this dataset, we include a very large number of covariates in $X_{i,j}$. See Table 4. They are complaint j -specific characteristics, including the (log-transformed) rank in the complaint sequence, the (log-transformed) number of days since the complaint $j - 1$ to j , the (log-transformed) average number of days across all past j complaints, whether a crash/fire was reported, whether there are injuries/deaths, whether the complaint was from one of the top 10 most populated states, whether the complaint was reported online, whether the complaint was about a car with an anti-brake system, the (log-transformed) number of months from the launch of MMY i to the creation of complaint j . We also include many different fixed effects to capture the characteristics of MMY i .

We randomly split all of the MMYs into in-sample (80%) and out-of-sample (20%) data sets to evaluate model performance. We further randomly split the in-sample MMYs into a training data set (approximately 90%) and a tuning data set (approximately 10%) to determine the model hyperparameters.⁷ As a result, our training/tuning/evaluation data set has 1,185/170/338 MMYs (894/129/255 of which were recalled) and 44,735/4,349/13,269 consumer complaints. Using our proposed SK metamodel, we find that the optimal hyper-parameters $[\alpha^*, \gamma^*, \sigma^{2*}, a^*, b^*, K^*]$ is $[0.20, 0.019, 0.27, 0.02, 2.10, 23]$. The details are provided in Web Appendix D.

5 Substantive Outputs

In this section, we report the substantive outputs from the corpus, which show very good face validity. These observations confirm that the HDPYP can provide reliable insights for manufacturers and

⁷Note that a recall statement may be used for multiple MMYs that pertain to a similar defect issue. Therefore, in the data partitioning process, we grouped MMYs that belong to the same recall statement and these correlated MMYs can only appear in either in-sample or out-sample data sets (but not both).

regulators. Importantly, as mentioned earlier, these insights would not be possible with a purely descriptive modeling approach to analyze consumer complaints.

5.1 Topics and Defect Components

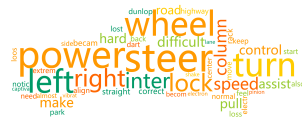
We first report the extracted topics. To facilitate understanding, we follow the literature to present the 50 most frequent words in each of the topics, using word clouds in which font size of each word is proportional to its frequency within the respective topic. All the 23 word clouds are reported in Figures 5. On the basis of these most relevant words, we label the topics “Steering and wheels,” “Suspension,” “Tires and wheels,” “Service brake,” “Engine related issues,” “Warranty,” “Hybrid propulsion system and trailer hitches,” “Lighting,” “Road condition,” “Airbag,” “Parking brake and forward collision avoidance,” “Multiple communication and feedback,” “Fuel system,” “Structure”, “Visibility,” “Equipment and air condition”, “Power train,” “Traction control system,” “Back over prevention,” “Electrical system related issues,” “Vehicle speed control,” “Seat and seatbelt,” and “Latches, locks, and linkages.”

We have a few observations regarding the extracted topics. First, these 23 topics cover all possible 26 recall components by the NHTSA.⁸ Eight out of the 23 topics enhance the clarity of defect explanations by focusing on two related components. For example, topic 1 combines “steering” and “wheels”, while topic 22 combines “seat” and “seat belt”. The detailed correspondence between our extracted topics and specific components are shown in Table 3. Second, consumers also raise concerns regarding the state of the vehicle (Topic 6), road conditions (Topic 9), and the communication and feedback processes with the manufacturer (Topic 12) in their complaints. Lastly, the most relevant words rarely appear repeatedly across different topics, except for certain exceptional cases where specific words may be associated with multiple defective issues (e.g., wheel). In sum, the extract topics are interpretable, coherent, and consistent with our expectations regarding possible vehicle defects.

5.2 Document Topic Distributions

The HDPYP can estimate the topic distribution of each document. For interested readers, in Table W3 of Web Appendix E, we report a few sample recall statements, along with their estimated topic distributions (i.e., $v_{i,j}$), and the expected topic distributions of its resulting recall statement based on all of its associated consumer complaints (i.e., θ_j). Overall, we find that the estimations based on the HDPYP are both intuitive and consistent.

⁸Note that two components were rarely discussed in consumer complaints nor in the resulting recall statements: *communication* (appearing in 0.2% of complaints) and *lane departure* (appearing in 0.02% of complaints).



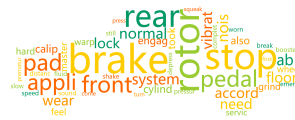
(a) Topic 1: "Steering and wheels"



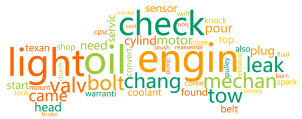
(b) Topic 2: "Suspension"



(c) Topic 3: "Tires and wheels"



(d) Topic 4: "Service brake"



(e) Topic 5: "Engine related issues"



(f) Topic 6: "Warranty"



(g) Topic 7: "Hybrid propulsion system and trailer hitches"



(h) Topic 8: "Lighting"



(i) Topic 9: "Road condition"



(j) Topic 10: "Airbag"



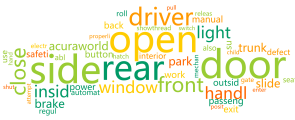
(k) Topic 11: "Parking brake and forward collision avoidance"



(l) Topic 12: "Multiple communication and feedback"



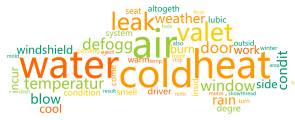
(m) Topic 13: "Fuel system"



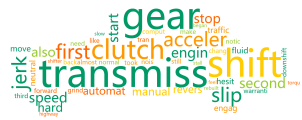
(n) Topic 14: "Structure"



(o) Topic 15: "Visibility"



(p) Topic 16: "Equipment and air condition"



(q) Topic 17: "Power train"



(r) Topic 18: "Traction control system"



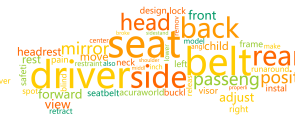
(s) Topic 19: "Back over prevention"



(t) Topic 20: "Electrical system related issues"



(u) Topic 21: "Vehicle speed control"



(v) Topic 22: "Seat and seat belt"



(w) Topic 23: "Latches, locks and linkages"

Figure 5: (Color online) Word Clouds of Extracted Topics by the HDPYP

Table 3: Labeled Topics and Defect Components

Topic	Topic label	#Components	Components
Topic1	steering and wheels	2	steering; wheels
Topic2	suspension	1	suspension
Topic3	tires and wheels	2	tires; wheels
Topic4	service brake	1	service brake
Topic5	engine related issues	2	engine; engine and engine cooling
Topic6	warranty	—	—
Topic7	hybrid propulsion system and trailer hitches	2	hybrid propulsion system; trailer hitches
Topic8	lighting	2	exterior lighting; interior lighting
Topic9	road condition	—	—
Topic10	air bags	1	airbag
Topic11	parking brake and forward collision avoidance	2	parking brake; forward collision avoidance
Topic12	multiple communication and feedback	—	—
Topic13	fuel system	1	fuel system
Topic14	structure	1	structure
Topic15	visibility	1	visibility
Topic16	equipment and air condition	1	equipment_air condition
Topic17	power train	1	power train
Topic18	traction control system	1	traction control system
Topic19	back over prevention	1	back over prevention
Topic20	electrical system related issues	2	electrical system; electronic stability control
Topic21	vehicle speed control	1	vehicle speed control
Topic22	seat and seatbelt	2	seat and seatbelt
Topic23	latches/locks/linkages	1	latches/locks/linkages

Note. “—” represents not included in complaint component list.

Next, we formally test whether the HDPYP can detect a meaningful relationship between complaints and their associated recall statements. For each recalled MMY i , we expect that the topic distribution of its recall statement, θ_i , is more similar to the topic distributions of its associated complaints, $v_{i,j}$, than to that of the non-associated complaints. We construct the non-associated complaints for each recalled MMY i as a set of complaints that are about the same automaker as MMY i but are not associated with any recall in our data. We find confirming evidence across all of the automakers that the estimated topic distributions of the recall statements are indeed significantly (about 30%) more similar to the topic distributions of their associated complaints than to that of the non-associated complaints ($p < 0.05$). The details are reported in Table W4 of Web Appendix E.

In Section 4, we show that there are minimal differences between the recalled and the non-recalled complaints in their metadata. We now examine whether they can exhibit meaningful differences in their topic distributions estimated by the HDPYP. Figure 6 compares the average topic distributions of recalled and non-recalled complaints. It is evident that recalled complaints, as a collective, exhibit significantly higher proportions in relation to component-related topics ($p < 0.001$) than non-recalled complaints, such as Topic 1 (steering and wheels), Topic 3 (tires and wheels), and Topic 8 (lighting). In addition, non-recalled complaints tend to place greater emphasis on non-safety concerns ($p < 0.001$), such as

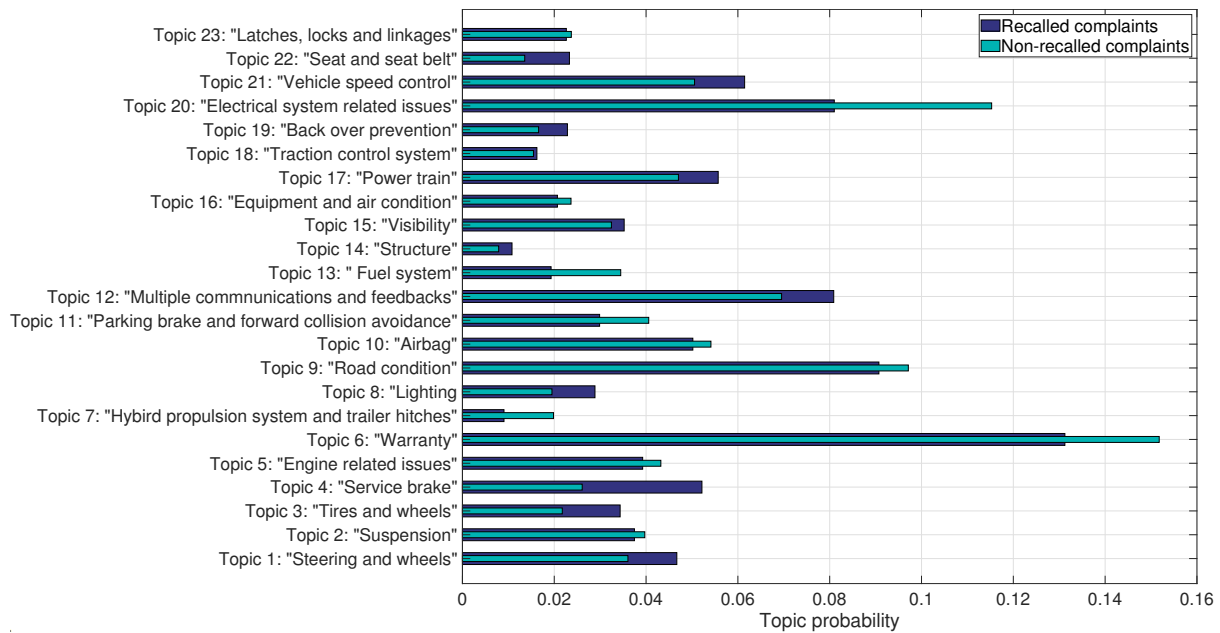


Figure 6: (Color online) Average Topic Distributions of the Recalled Complaints versus the Non-recalled Complaints

Topic 6 (warranty) and Topic 9 (road condition), while the recalled complaints have greater emphasis on non-component related issue ($p < 0.001$) like Topic 12 (multiple communications and feedbacks). A plausible rationale for this phenomenon lies in the persisting occurrence of consumer reports regarding a specific defect issue, which serves as evidence of the problem's sustained existence beyond sporadic instances. As a result, there arises the likelihood that this issue may present a safety hazard capable of adversely affecting normal driving, thereby attracting the scrutiny of regulatory entities. In sum, our findings further confirm the importance of properly mining the textual information in consumer complaints for recall management.⁹

5.3 What Complaints are Important

We can derive interpretable insights on what kinds of consumer complaints are more or less important based on the metadata, using the estimated λ in the HDPYP. The estimation results are shown in Table 4. We find that most of the covariates are statistically significant. For example, complaints that are posted either earlier in the sequence or online tend to have a higher importance weight ($p < 0.001$). The time gap between the current and the previous complaint is negatively associated with the importance of the current complaint ($p < 0.001$). Interestingly, *crash*, *fire* and *injured* have positive effects on

⁹Note that although one can also estimate the topic distribution of each consumer complaint using a simple LDA and then compare the differences between recalled and non-recalled complaints, we will show later in Section 6.2 the estimated topic distributions from the LDA are much less meaningful than these obtained from the HDPYP, and hence fail to capture subtle differences across documents.

importance weights ($p < 0.01$), whereas *deaths* does not exhibit a significant impact to the same extent. We find that this is probably driven by the rarity of death in this context (0.2% of all of the complaints) and the high correlation (0.90) between a crash and death. Our finding also kind of echoes with prior research (e.g., Astvansh et al., 2022) that the number of deaths is insignificantly associated with both voluntary and involuntary recalls. These MMY-specific variables control for the overall variations in the importance weights across complaints. Higher coefficients indicate smaller variations. We find that complaints about certain brands (e.g., Buick), body types (e.g., convertibles), cylinders (e.g., V8), and drive types (e.g., 4WD) will result in relatively small variations across complaints.

5.4 Complaint Importance

We finally report the estimated importance weight of each complaint j to its associated recall statement i , i.e., $\rho_{i,j}$. We compute the variation of $\rho_{i,j}$ using its coefficient of variation (CV) across all of the complaints associated with recall i . We find that the CV ranges from 16.64% to 1682.06%, and that more than 97% of the recalls have a CV that is greater than 50%. This finding confirms that only a small number of complaints, out of a large volume of complaints, actually contribute to the resulting recall decision. In addition, we test whether a consumer complaint with a higher importance weight is semantically more similar to its associated recall statement. We conduct a regression model of $\rho_{i,j}$ on $\text{Cos}(\theta_i, \mathbf{v}_{i,j})$, while controlling for the recall fixed effect and using a total of 29,514 observations. We find confirming evidence that the coefficient of $\rho_{i,j}$ is significantly positive (0.467, $p < 0.001$).

6 Validating Predicted Complaint Importances

In this section, we further validate the predicted importance of each consumer complaint from the HDPYP through two applications. We demonstrate how they could be used in practice for recall management as regulators and manufacturers receive consumer complaints over time.

6.1 Application of Predicting Recall Incidence/Components

One primary application is to assess the likelihood of a recall as new complaints surface over time, ultimately minimizing the adverse effects associated with product defects (Aragon et al., 2019). We construct a panel data at the MMY-month level for this application. We consider four widely recognized predictive models: XGBoost (Chen and Guestrin, 2016), logistic regression, the random forest method,

Table 4: Estimated λ

Metadata	Estimation	Metadata	Estimation
(Intercept)	-15.58 (2.21) ***	4WD	1.72 (0.52) **
log(CompIndex)	-11.70 (0.17) ***	RWD	-0.27 (0.31)
Crash	1.68 (0.22) ***	Multiple types	
Fire	1.91 (0.36) ***	Cylinder Type	
Injured	1.01 (0.26) ***	EV	-0.13 (1.51)
Death	2.84 (1.24) *	I3	15.80(4.94) **
Top10State	0.73 (0.1) ***	I4	-0.09 (0.17)
CompType	4.34 (0.12) ***	I6	-0.51 (0.65)
AntiBrakes	0.44 (0.15) **	V10	5.45 (3.58)
log(TimeSinceLaunch)	-2.29 (0.60) ***	V6	-0.80 (0.18) ***
log(TimeSinceLaunch2)	0.35 (0.23)	V8	1.17 (0.31) ***
log(TimeGapFromPreComp)	0.3 (0.10) **	Flat4	-2.84 (0.81) ***
log(AveTimeGapComps)	2.76 (6.81)	Rrotary0	0.39 (1.38)
Automaker		Multiple types	
Acura	2.32 (0.53) ***	Transmission Type	
BMW	-0.01 (0.71)	Automatic	0.34 (0.18)
Buick	2.57 (0.62) ***	Manual	0.49 (0.22) *
Cadillac	-2.48 (0.70) ***	Multiple types	
Chevrolet	-1.72 (0.36) ***	Engine Type	
Chrysler	0.27 (0.49)	Electric	5.11 (2.16) *
Dodge	-1.31 (0.43) **	Flex-fuel	3.46 (0.61) ***
Ford	-1.41 (0.34) ***	Gas	-0.39 (0.19) *
GMC	-0.87 (0.63)	Hybrid	-0.15 (0.40)
Honda	-0.60 (0.39)	Multiple types	
Hyundai	0.33 (0.39)	Model Year	
Jeep	-1.06 (0.46) *	2000	2.30 (0.59) ***
Kia	-0.76 (0.37) *	2001	0.01 (0.61)
Lexus	-0.95 (0.49)	2002	1.27 (0.59) *
Mazda	-2.19 (0.4) ***	2003	0.23 (0.59)
Mercedes-Benz	-0.52 (0.55)	2004	-0.62 (0.59)
Nissan	-0.86 (0.39) *	2005	0.19 (0.59)
Subaru	-0.64 (0.94)	2006	-0.12 (0.60)
Toyota	-0.51 (0.35)	2007	-1.17 (0.57) *
Volkswagen		2008	-0.40 (0.59)
Body Type		2009	0.55 (0.62)
Convertible	4.62 (0.76) ***	2010	-0.05 (0.58)
Coupe	-0.01 (0.63)	2011	1.21 (0.60) *
Hatchback	1.66 (0.26) ***	2012	-0.16 (0.59)
Minivan	2.75 (0.3) ***	2013	0.03 (0.58)
Sedan	1.07 (0.26) ***	2014	1.39 (0.62) *
SUV	0.61 (0.18) **	2015	0.99 (0.63)
Wagon	0.97 (0.36) **	2016	3.07 (0.69) ***
Multiple types		2017	3.64 (0.70) ***
Drive Type		2018	0.58 (0.65)
AWD	0.89 (0.53)	2019	
FWD	-0.27 (0.23)		

Note. The category without the estimated coefficient is set as the baseline category for each variables. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

and the extended Cox model.¹⁰ Each model is calibrated using in-sample training MMYs, with its accuracy is evaluated using out-sample MMYs. Calibration of the prediction thresholds was performed independently, using the in-sample tuning MMYs (Naumzik et al., 2022).

Following the industry practice described in Stout (2019), we let the input of these prediction models for each MMY i at time t contain a large number of fixed effects, capturing the characteristics of the MMY i (see Table 4), and rich characteristics of its complaints. They are the cumulative number of complaints, the average and cumulative amount of metadata across all of the complaints (see Table 2), the number of months since MMY i 's launch, and the average topic distributions of the complaints.

6.1.1 Benchmarks

One key question is how to aggregate the topic distributions across complaints. The existing literature and our study both have shown that the severity of the reported issues varies greatly across complaints and a recall is triggered by only a few complaints. The HDPYP can capture such a reality effectively, and the actual implementation is described as follows. First, using the posterior estimates of all of the parameters in the HDPYP, the topic distribution of each newly arrived consumer complaint can be quickly estimated. Second, the relative importance weights of all of the consumer complaints are updated using the estimated Dirichlet-multinomial regression model. Third, the weighted average topic distribution across all of the complaints are updated using $\hat{\theta}_i = \sum_j \hat{\rho}_{i,j} \hat{\mathbf{v}}_{i,j}$. Lastly, the likelihood of recall is predicted using a trained prediction model with these updated covariates as inputs. We refer to this approach of aggregating the textual information in consumer complaints by HDPYP _{ρ} .

To measure the incremental gain of our proposed approach, we compare its prediction accuracy with that of four benchmark inputs. The first (labeled by HDPYP_{Equal}) uses the topic distributions of the consumer complaints estimated by the HDPYP, assuming that they contribute equally. The second benchmark (labeled by LDA_{Equal}) uses the topic distributions of the consumer complaints estimated by the LDA, assuming that they contribute equally.¹¹ The third benchmark (labeled by Component) does not include the topic distributions of any consumer complaints, but solely incorporates component information provided by consumers. The last benchmark (labeled by n.a.) excludes both the topic distributions of any consumer complaints and the component information altogether. This is used to

¹⁰The extended Cox model augments the conventional Cox model for survival analyses, specifically accommodating time-varying covariates (Bentley and Friesner, 2020).

¹¹The LDA is trained by treating complaints and recall statements as independent documents. We optimize the hyperparameters of the LDA using our proposed SK metamodeling method; and we find that it also has 23 topics which have very similar labels as these extracted by the HDPYP.

understand the general value of incorporating textual information.¹²

6.1.2 Predicting Recall Incidence in Advance

We first consider how well one can predict an impending recall incidence in advance (among these recalled MMYs in the out-sample data set). Such foresight is crucial for stakeholders to take preparatory actions, thereby circumventing extended damages resulting from belated recalls. The out-of-sample AUC for predictions made several months in advance are compared in Table 5 for different prediction models and inputs. Notably, we find that the HDPYP ρ -based model inputs consistently outperform other input methods across all prediction models ($p < 0.05$). Specifically, under the XGBoost predictive model with up-to one-year horizon, the HDPYP ρ -based inputs achieve an AUC of 80%-86%, a 14% improvement over the n.a.-based inputs. As expected, using equally weighted topic distributions from LDA and HDPYP did not significantly improve predictability compared with not using complaints at all (i.e., the n.a. inputs), under random forest and extended Cox models. Across all prediction models, using component variables leads to either no or minimal improvement compared with the n.a. inputs. These findings further confirm the effectiveness of encoding weighted information from the HDPYP in recall management. Intuitively, as the forecast horizon extends, the predictive power of a model with a specific input specification decreases. This suggests the importance of timely mining consumer complaints to maintain prediction accuracy. Finally, XGBoost performs the best in predicting recalls across the four prediction models. This finding is consistent with prior prediction tasks in marketing contexts (e.g., Rafieian and Yoganarasimhan, 2021; Zhang and Luo, 2023).

For completeness, we also evaluate the out-of-sample accuracy of predicting whether an MMY should be recalled now given all the information observed so far. There we consider five different performance metrics. The results are reported in Table W6 of Web Appendix F. We find that the overall findings are consistent with the accuracy of predicting in advance.

6.1.3 Predicting Crippling Recall Events in Advance

To further elucidate the value of enhanced accuracy in recall prediction endeavors, we spotlight three financially devastating recall incidents within the automotive sector observed in our data set.¹³ These

¹²Note that we also consider weighting all of the covariates using ρ . However, we find that this approach leads to significantly lower prediction accuracy than the approach of only weighting the topic distributions of complaints, though HDPYP ρ is still significantly better than the other benchmarks on most performance metrics. This finding is intuitive, given that ρ in the HDPYP only captures the semantic relationships between the two types of documents, whereas all of the metadata only serve as the input in ρ (rather than ρ itself).

¹³See: <https://autoly.io/6-expensive-automotive-recalls-that-completely-shook-the-industry/>.

Table 5: Predicting Recall Incidence Early in Advance (%)

Model	Input	1 month ahead	3 months ahead	6 months ahead	12 months ahead
XGBoost	HDPYP _{ρ}	85.89*	83.89*	81.40*	79.74*
	HDPYP _{Equal}	80.25	78.59	77.13	74.29
	LDA _{Equal}	79.43	78.58	76.58	74.78
	Component	76.55	74.81	73.73	72.16
	n.a.	75.06	74.35	71.84	71.33
Logistic Regression	HDPYP _{ρ}	79.19*	78.42*	77.30*	76.28*
	HDPYP _{Equal}	75.25	74.90	74.08	72.79
	LDA _{Equal}	75.80	73.63	72.27	70.08
	Component	74.67	73.88	72.47	72.93
	n.a.	72.38	70.04	69.85	67.79
Random Forest	HDPYP _{ρ}	74.93*	71.71*	67.66*	65.17*
	HDPYP _{Equal}	64.27	64.11	63.14	58.84
	LDA _{Equal}	64.69	63.07	62.24	56.86
	Component	65.44	63.02	62.90	61.98
	n.a.	66.47	63.76	61.67	60.56
Extended Cox	HDPYP _{ρ}	66.88*	64.38*	62.70*	60.63*
	HDPYP _{Equal}	59.24	57.64	55.45	55.08
	LDA _{Equal}	59.46	57.65	56.06	53.19
	Component	58.71	56.41	54.14	52.39
	n.a.	59.36	58.47	56.23	55.96

Note. The table reports out-of-sample AUC values in percentages. These input methods differ in how the textual information in consumer complaints is processed and aggregated to construct the covariates that enter a given prediction model. *The input method is the best at $p < 0.05$.

recall events are: Toyota accelerator pedals (Case 1), which involved eight affected MMYs and caused \$2 billion financial loss and 40 deaths; General Motors ignition switches (Case 2), which involved five affected MMYs and caused \$3 billion loss and 124 deaths; and Volkswagen diesel engine (Case 3), which involved 11 million affected vehicles and caused \$18.3 billion loss. For each of these MMYs, we predict its recall probability up-to 24 months in advance, leveraging HDPYP _{ρ} -based inputs within the XGBoost framework. Here we only consider the n.a.-based inputs as a benchmark because Table 5 suggests that its accuracy is tied with the LDA_{Equal}-based inputs in most cases.

The predicted recall probability per MMY under different lead times are compared in Table 6. In the left panel, we find that the HDPYP _{ρ} -based inputs yield consistently high recall probability for all of these affected MMYs, all above 0.65, even when predicting 12 months in advance. Given the modest recall threshold (below 0.30 for XGBoost), our proposed model would have advocated recalls for all these MMYs a substantial 12 months ahead of their actualized recall dates. Such proactive measures could have substantially mitigated corporate financial hemorrhages and, more imperatively, safeguarded consumers from ensuing property damage or tragic fatalities. In comparison, in the right panel, we

Table 6: Predicted Probability of Recall Early in Advance under the XGBoost Model

Affected MMY	Lifetime (months)	HDPYP _ρ -based Inputs				n.a.-based Inputs			
		3 months ahead	6 months ahead	12 months ahead	24 months ahead	3 months ahead	6 months ahead	12 months ahead	24 months ahead
Toyota 4runner 2004	84.03	0.777	0.736	0.741	0.724	0.200	0.180	0.185	0.183
Toyota Avalon 2007	30.60	0.755	0.714	0.702	0.719	0.692	0.662	0.651	0.668
Toyota Avalon 2008	18.43	0.715	0.710	0.704	-	0.618	0.584	0.594	-
Toyota Camry 2009	6.23	0.738	0.716	-	-	0.676	0.641	-	-
Toyota Camry Hybrid 2007	30.60	0.702	0.669	0.673	0.680	0.367	0.299	0.314	0.304
Toyota Highlander 2005	15.60	0.759	0.741	0.714	-	0.68	0.678	0.671	-
Toyota Prius 2007	30.60	0.839	0.818	0.822	0.820	0.777	0.785	0.760	0.766
Toyota Prius 2008	18.43	0.773	0.729	0.742	-	0.687	0.683	0.676	-
Toyota Sequoia 2008	22.03	0.692	0.651	0.680	-	0.414	0.403	0.407	-
Toyota Venza 2009	10.03	0.732	0.710	-	-	0.590	0.565	-	-
Chevrolet Camaro 2011	39.20	0.767	0.764	0.743	0.739	0.625	0.639	0.624	0.630
Chevrolet HHR 2010	47.03	0.826	0.827	0.819	0.824	0.246	0.227	0.238	0.255
Chevrolet HHR 2011	34.87	0.742	0.686	0.688	0.704	0.632	0.59	0.601	0.618
Chevrolet Impala 2013	14.93	0.733	0.697	0.694	-	0.614	0.593	0.636	-
Chevrolet Tahoe 2012	33.47	0.718	0.705	0.692	0.687	0.246	0.188	0.179	0.174
Volkswagen Golf 2010	18.33	0.755	0.697	0.706	-	0.407	0.381	0.373	-
Volkswagen Golf 2011	6.17	0.778	0.751	-	-	0.720	0.687	-	-

Note. The second column denotes the number of months between the release date of an MMY and the recall date of the MMY.

find that the n.a.-based inputs produce relatively small recall probability in most cases. As a result, even when projecting only a few months ahead, the reliance on the n.a.-based inputs translates to an omission in recommending recalls for several MMYs (e.g., Toyota 4runner 2004, Toyota Sequoia 2008, and Chevrolet HHR 2010).

6.1.4 Predicting the Defect Components in Advance

The calibrated HDPYP can also be used to provide insights into the specific components of an impending recall event. To illustrate this, we conduct an out-sample prediction of the actual defect components mentioned in recall statements. In our empirical context, we find that over 99% of consumers select up-to three possible defect components. Thus, we consider the prediction to be accurate if any of the top three components predicted by a given method matches with the defect component(s) mentioned in the associated recalled statement. We consider four benchmark prediction methods. The first utilizes the topic distributions of the top 10 complaints ranked by the HDPYP (referred to as HDPYP_{Top10}). The second and third methods use the topic distributions of all of the complaints estimated by HDPYP and LDA (referred to as HDPYP_{All} and LDA_{All}, respectively). The last method exclusively relies on component choices directly reported by consumers (referred to as Component_{All}). For the first three methods (using topic modeling), we identify the top three topics (and their components) based on the estimated topic distribution of each complaint and then identify the top three components across all of the complaints associated with the same MMY.

Table 7: Predicting Recalled Components in Advance

Input Method	0 month ahead	1 month ahead	3 months ahead	6 months ahead
HDPYP _{Top10}	0.271*	0.283*	0.270*	0.270*
HDPYP _{All}	0.247	0.256	0.254	0.251
LDA _{All}	0.114	0.110	0.115	0.119
Component _{All}	0.212	0.224	0.221	0.224

Note. * indicates the best input method for the out-sample dataset ($p < 0.05$).

Table 7 compares the prediction accuracy across methods under different lead times. We find that the HDPYP_{Top10} significantly outperforms the other methods across all time frames ($p < 0.05$), while the LDA_{All} performs the worst ($p < 0.05$). The accuracy under HDPYP_{Top10} is increased by about 24% compared with directly using consumer self-reported defect components. This further confirms the value of focusing on the few most important complaints for recall management. In conclusion, we demonstrate the practical value of estimating the importance of consumer complaints using our proposed HDPYP,

particularly when applied to the realm of recall management.

6.2 Application of Drafting Recall Statements

When practitioners anticipate an impending recall event, it is timely important to draft a recall statement, and regulators must verify whether this statement reflects the issues raised in consumer complaints. In this section, we showcase how the HDPYP can be applied to understand which complaints and topics are worthy of summary, and verify to what extent a recall statement aligns with consumer complaints.

6.2.1 Predicting the Topic Distributions of Recall Statements

We start by examining how well one can predict the topic distribution of a resulting recall statement based on the observed consumer complaints so far. We compare the predictions based on HDPYP_ρ , $\text{HDPYP}_{\text{Equal}}$, $\text{LDA}_{\text{Equal}}$, and $\text{LDA}_{\text{Pooled}}$. The last benchmark pools complaints associated with the same MMY and estimates its distribution using the LDA. We use the commonly used metric, *perplexity*, for measuring how well the estimated topic distribution fits the actual content of a recall statement (Liu and Toubia, 2018; Toubia, 2020). Table 8 compares perplexity across different methods for both the in- and out-sample data sets. We find that the HDPYP_ρ performs significantly better than all of the other benchmarks ($p < 0.05$), improving the perplexity by about 40%. This provides evidence that the HDPYP can effectively identify important complaints and possible defect components for recall management.

Table 8: Perplexity Scores of Predicted Recall Statements

	In-sample	Out-sample
HDPYP_ρ	725.488*	737.239*
$\text{HDPYP}_{\text{Equal}}$	829.913	878.480
$\text{LDA}_{\text{Equal}}$	1181.118	1176.145
$\text{LDA}_{\text{Pooled}}$	1194.135	1261.070

Note. * indicates the best method at $p < 0.05$.

6.2.2 Automating the Generation of Recall Statements

The HDPYP can also be combined with pre-trained LLMs (e.g., BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), and T5 (Raffel et al., 2020)) to create the actual content of recall statements. In this study, we choose T5 for two reasons. First, T5 is an open-source project with remarkable transparency in its training process and model weights (Raffel et al., 2020). This transparency makes it a versatile choice applicable to a wide range of domains and applications, all without incurring extra licensing expenses.

Second, T5 demonstrates exceptional generalization abilities, especially in text-to-text tasks. Empirical evidence shows that it achieves state-of-the-art performance when fine-tuned on diverse, domain-specific datasets, surpassing open-source models like BERT and GPT-2 (Kale and Rastogi, 2020). Therefore, T5 is more suitable for our research objective.

The flow of using T5 to generate recall statements is depicted in Figure W3 of Web Appendix G.¹⁴ We adapt T5 to our research domain using the in-sample complaint-recall data set with the specific task of summarizing the recall statements. To this end, we provide a consistent input framing that allows T5 to recognize our task (i.e., generating a recall statement) that should be performed given an input sequence of tokens (i.e., formatted complaint-recall pairs). Thus, the role of the HDPYP in this process is to identify the top few most important consumer complaints that will be used as input to fine-tune T5 so that the generated recall statements are relevant.

We compare the auto-generated recall statements across the following input methods: (1) the top 10 complaints predicted by the HDPYP; (2) 10 randomly selected complaints; (3) only complaints that are associated with fires, crashes, injuries, or deaths; and (4) all of the complaints. We employed a diverse set of evaluation metrics for generated content, including ROUGE (Lin, 2004), METEOR (Lavie and Agarwal, 2007), MoverScore (Zhao et al., 2019), BertScore (Zhang et al., 2020). ROUGE assesses text overlap between generated and reference text, encompassing unigram (ROUGE₁), bigram (ROUGE₂), and the longest common word sequence (ROUGE_L). METEOR extends this focus by considering word stemming, synonyms, and word order. In contrast, MoverScore and BertScore emphasize semantic similarity. MoverScore calculates Word Mover's Distance (Kusner et al., 2015), while BertScore uses a pre-trained BERT model to compute cosine similarity between text embeddings. For all of these metrics, higher scores indicate better content quality.

Table 9 presents the evaluation results. We find that our HDPYP-aided input method outperforms all the benchmark inputs across all metrics ($p < 0.05$). This observation confirms that the HDPYP not only identifies important complaints effectively but also enhances the quality of auto-generated recall statements by LLMs. For those interested, we provide the auto-generated recall statements based on these input methods in Tables 10-11, along with two additional examples in Web Appendix G.

¹⁴Note that such generative AIs are trained using available large-scale data without information on specific problem context, so to extend their functionality, these models usually need to be combined with other input models by fine-tuning prompts that are phrased as instructions. To trade between accuracy and efficiency, we apply the T5_{base} architecture, which is characterized by 12 blocks for encoders and decoders and is pre-trained on a large-scale general-purpose corpus (Raffel et al., 2020). We use *Python* packages to fine-tune the T5_{base} model, employing an early stopping strategy with a learning rate of 0.001 and a maximum of 15 epochs.

Table 9: Comparing the Auto-Generated Recall Statements

Input Method	ROUGE ₁	ROUGE ₂	ROUGE _L	METEOR	MoverScore	BertScore
Top-10 complaints from the HDPYP	0.3598*	0.2131*	0.3105*	0.2870*	0.3064*	0.4295*
Random 10 complaints	0.2894	0.1213	0.2362	0.2264	0.2514	0.3514
Complaints with fires, crashes, injuries, or deaths	0.2779	0.1207	0.2294	0.2266	0.2535	0.3520
All complaints	0.3122	0.1434	0.2572	0.2498	0.2742	0.3748

Note. * indicates the best input method for generating recall statements (by LLM T5) for the out-sample MMYs ($p < 0.05$).

Table 10: Auto-Generated Recall Statements Using Different Inputs - I

MMY: Ford Escort 2000	Recall campaign: 00V228001	Components: Seat belts
Recall defect and consequence summary:		
Vehicle description: passenger vehicles, light duty trucks, sport utility vehicles, and mini vans equipped with TRW seat belt buckle assemblies. The buckle base of these seat belt assemblies were not properly heat treated and therefore do not pass the load bearing requirement of federal motor vehicle safety standard No. 209, "seat belt assemblies." in the event of a vehicle crash, the occupant may not be properly restrained.		
Top-10 complaints ranked by HDPYP:		
Vehicle description: passenger vehicles equipped with TRW seat belt buckle assemblies. The buckle base of these seat belt assemblies were not properly heat treated and therefore do not pass the load bearing requirement of federal motor vehicle safety standard No. 209, "seat belt assemblies." in the event of a vehicle crash, the occupant may not be properly restrained.		
Random 10 complaints:		
On certain passenger vehicles, the fuel pump may corrode and separate from the motor housing. This could result in corrosion of the fuel pump housing which could lead to leakage of fuel. Loss of fuel may increase the risk of a fire.		
Complaints with fired, crashed, injured or death:		
On certain passenger vehicles, the rear suspension struts may not have been properly assembled. The rear struts could become loose or loose due to an improper welding process. This could result in the front struts becoming loose and thus increasing the risk of a crash.		
All complaints:		
Vehicle description: passenger vehicles. Some of these vehicles fail to comply with the requirements of federal motor vehicle safety standard (FMVSS) number 108, "occupant crash protection." In the event of a crash necessitating deployment of an air bag, the driver may not be properly restrained.		

7 Conclusion

Although there is a large stream of research on leveraging NLP tools to analyze UGC, consumer complaints remain largely underexplored. We bridge this gap by introducing an efficient and scalable topic model, the HDPYP, for extracting valuable features from consumer complaints and related recall statements. The HDPYP can extract defect topics, the topic distribution of a complaint, the significance of complaints, and identify pivotal complaint types. To facilitate an efficient and scalable model inference, we also develop a blocked and collapsed Gibbs sampling and a new method of searching for the optimal model hyperparameters.

In our empirical study focusing on the U.S. automobile industry, we demonstrate that the HDPYP can extract meaningful substantive insights from the corpus, subsequently aiding recall management

Table 11: Auto-Generated Recall Statements Using Different Inputs - II

MMY: Volkswagen CC 2013	Recall campaign: 15V483000	Components: Air bags
Recall defect and consequence summary:		
Volkswagen group of America, inc. (Volkswagen) is recalling certain model year 2010-2014 Volkswagen CC, and Passat, 2010-2013 EOS, 2011-2014 Golf, Gti, Jetta, and Tiguan, and 2012-2014 Jetta Sportwagen vehicles. In the affected vehicles, debris may contaminate the air bag clock spring, a spiral wound, flat cable that keeps the air bag powered while the steering wheel is being turned. This contamination may tear the cable and result in a loss of electrical connection to the driver's frontal air bag. A loss of electrical connection to the driver's frontal air bag will prevent the air bag from deploying in the event of a vehicle crash, increasing the risk of injury.		
Top-10 complaints ranked by HDPYP:		
Volkswagen group of America, inc. (Volkswagen) is recalling certain model year 2013-2014 Volkswagen CC, and Passat, 2013-2014 Golf, Gti, Jetta, and Tiguan, and 2012-2014 Jetta Sportwagen vehicles. In the affected vehicles, debris may contaminate the air bag clock spring, a spiral wound, flat cable that keeps the air bag powered while the steering wheel is being turned.		
Random 10 complaints:		
Volkswagen group of America, inc. (Volkswagen) is recalling certain model year 2013-2014 CC vehicles manufactured January 1, 2013, to December 31, 2013. The affected vehicles may have been assembled with improperly torqued bolts which could allow the bolts to separate from the bolts. This condition can cause the bolts to separate from the bolts and prevent the bolts from being able to be moved out of position.		
Complaints with fired, crashed, injured or death:		
On certain passenger vehicles, the fuel pump hose may not have been properly installed. Fuel leakage in the presence of an ignition source could result in a fire.		
All complaints:		
Volkswagen is recalling certain model year 2013 CC vehicles manufactured from January 1, 2012, through December 31, 2013. the vehicle may have been equipped with an improperly installed brake booster. This could cause the engine to stall without prior warning and increase the risk of a crash.		

decisions, such as identifying defect components, initiating a recall, and drafting/verifying recall statements. We emphasize that the potential applications of the HDPYP are vast and tailored to decision-makers needs. This model offers manufacturers and regulators a preemptive alert system, allowing for real-time consumer complaint monitoring, which in turn minimizes business losses and safeguards consumers from potential harm or loss. For manufacturers, it is a powerful tool to pinpoint products that demand immediate attention, gauge recall risks, identify potential defect themes, and even predict recall statement content, all with minimal manual oversight. Regulators benefit too. The HDPYP framework shines a spotlight on crucial complaints that trigger recalls and bolsters their investigative processes. Furthermore, when drafting recall statements, our approach empowers regulators to measure the alignment between firm-submitted notices and the actual issues highlighted by consumers. Hence, our proposed method not only liberates regulators from the tedious task of monitoring but also accelerates their market responses, ensuring safety without significant budget strains.

Broadly, our research underscores the need to move past traditional measures, such as volume and valence, when scrutinizing consumer complaints. Our findings indicate that the discrepancies in volume- and valence-related metrics between already-recalled and yet-to-be-recalled MMYs are marginal or statistically negligible. Given that companies and regulators are investing considerable time and resources

in evolving technologies to ensure product safety, our model can provide valuable insights, potentially shaping their technological investment decisions.

The HDPYP introduces a novel semantic structure intrinsic to hierarchical topic modeling and hence contributes to the literature. The HDPYP can be applied in any context of documents in which multiple documents contribute to the creation of a summarizing document. An intriguing avenue for future research could involve adapting the HDPYP to sharing economy platforms known for their bilateral reviews. In such scenarios, our model has the potential to pinpoint pivotal historical reviews for both hosts and guests, which can influence transaction completions. This intelligence can guide platforms in making informed decisions about which sellers or reviews to present to potential buyers.

Our research has several limitations that can be explored in future research. First, because of data availability issues, we restrict our application to vehicle recalls. However, our proposed method can be similarly applied in other recall-intensive sectors, such as food and pharmaceuticals. Second, the collection of consumer complaints and the roles of regulators vary across industries. Thus, the actual value of the HDPYP in facilitating recall management for firms and regulators may vary according to context. Nevertheless, we demonstrate the potential value of the HDPYP to the automobile industry. Third, future research could incorporate potential recall dependence across automakers, as recent developments in the recall literature reveal that recalls might occur in clusters (Wowak et al., 2021). Lastly, future research could incorporate seeded topics in the HDPYP, which requires strong prior knowledge of domain-specific topics and further complicates the model inference algorithm.

Funding and Competing Interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. This work was supported by two funding institutions. The actual names and IDs are not provided for anonymity in the peer review process.

References

Ankenman, B. E., B. L. Nelson, and J. Staum (2010). Stochastic Kriging for simulation metamodeling. *Operations Research* 58(2), 371–382.

- Aragon, A., U. Huber, T. Möller, and H. Nick (2019). Return to sender: Resolving the automotive-recall resurgence. *McKinsey & Company*.
- Archak, N., A. Ghose, and P. G. Ipeirotis (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science* 57(8), 1485–1509.
- Astvansh, V., G. P. Ball, and M. Josefy (2022). The recall decision exposed: Automobile recall timing and process data set. *Manufacturing & Service Operations Management* 24(3), 1457–1473.
- Astvansh, V., Y.-Y. Wang, and W. Shi (2022). The effects of the news media on a firm’s voluntary product recalls. *Production and Operations Management* 31(11), 4223–4244.
- Bentley, J. P. and D. Friesner (2020). *Logistic Regression and Survival Analysis*. New York, NY: McGraw-Hill Education.
- Berger, J., A. Humphreys, S. Ludwig, W. W. Moe, O. Netzer, and D. A. Schweidel (2020). Uniting the tribes: Using text for marketing insight: Using text for marketing insight. *Journal of Marketing* 84(1), 1–25.
- Blei, D. M., T. L. Griffiths, and M. I. Jordan (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM* 57(2), 1–30.
- Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3(1), 993–1022.
- Buntine, W. and M. Hutter (2012). A Bayesian view of the Poisson-Dirichlet process. *ArXiv preprint arXiv: 1007.0296v2*.
- Büschken, J. and G. M. Allenby (2016). Sentence-based text analysis for customer reviews. *Marketing Science* 35(6), 953–975.
- Caron, F., W. Neiswanger, F. Wood, A. Doucet, and M. Davy (2017). Generalized Pólya urn for time-varying Pitman–Yor process. *Journal of Machine Learning Research* 18(27), 1–32.
- Che, X., H. Katayama, and P. Lee (2023). Product-harm crises and spillover effects: A case study of the volkswagen diesel emissions scandal in ebay used car auction markets. *Journal of Marketing Research* 60(2), 409–424.
- Chen, C., L. Du, and W. Buntine (2011). Sampling table configurations for the hierarchical Poisson-Dirichlet process. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis (Eds.), *Machine*

- Learning and Knowledge Discovery in Databases*, Berlin, Heidelberg, pp. 296–311. Springer Berlin Heidelberg.
- Chen, J., Y. Yang, and H. Liu (2021). Mining bilateral reviews for online transaction prediction: A relational topic modeling approach. *Information Systems Research* 32(2), 541–560.
- Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 785–794. Association for Computing Machinery.
- Chen, Y., S. Ganesan, and Y. Liu (2009). Does a firm's product-recall strategy affect its financial value? An examination of strategic alternatives during product-harm crises. *Journal of Marketing* 73(6), 214–226.
- Clifford Atiyef (2014). Recalls explained: How the government forces automakers to fix things. Available online at <https://www.caranddriver.com/news/a15362381/recalls-explained-how-the-government-forces-automakers-to-fix-things>.
- ConsumerReports.org (2015). The truth about car recalls. Available online at <https://www.consumerreports.org/cro/magazine/2015/04/the-truth-about-car-recalls/index.htm>.
- Coolidge, A. (2018). 'A child could die.' Honda Odyssey owners warned feds about seat dangers before Kyle Plush's death. Available online at <https://www.cincinnati.com/story/money/2018/12/18/honda-odyssey-owners-warned-feds-seat-dangers-before-kyle-plush-death/2115462002/>.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (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*, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics.
- DOT (2020). FY 2021 nhtsa congressional justification - 508 compliant. Available online at <https://www.transportation.gov/mission/budget/nhtsa-cj-fy-2021-estimates>.
- Du, L., W. Buntine, and H. Jin (2010). A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Machine Learning* 81(1), 5–19.
- Du, L., W. Buntine, H. Jin, and C. Chen (2012). Sequential latent Dirichlet allocation. *Knowledge and Information Systems* 31(3), 475–503.

- ECA Academy (2023). Complaint handling and recall management. Available online at <https://www.gmp-compliance.org/training/gmp-course-conference/complaint-handling-and-recall-management>.
- Fan, W., Z. Guo, N. Bouguila, and W. Hou (2021). Clustering-based online news topic detection and tracking through hierarchical Bayesian nonparametric models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York City, New York, USA, pp. 2126–2130.
- Fan, W., H. Sallay, and N. Bouguila (2017). Online learning of hierarchical Pitman–Yor process mixture of generalized Dirichlet distributions with feature selection. *IEEE Transactions on Neural Networks and Learning Systems* 28(9), 2048–2061.
- Foster, T., E. Makarova, and V. Telpis (2021). Making quality assurance smart. Available online at <https://www.mckinsey.com/industries/life-sciences/our-insights/making-quality-assurance-smart>.
- GAO (2016). Vehicle safety: Enhanced project management of new information technology could help improve nhtsa’s oversight of safety defects. Available online at the U.S. Government Accountability Office <https://www.gao.gov/products/gao-16-312>.
- Gao, Y., W. Duan, and H. Rui (2022). Does social media accelerate product recalls? Evidence from the pharmaceutical industry. *Information Systems Research* 33(3), 954–977.
- Hora, M., H. Bapuji, and A. V. Roth (2011). Safety hazard and time to recall: The role of recall strategy, product defect type, and supply chain player in the U.S. toy industry. *Journal of Operations Management* 29(7–8), 766–777.
- John O’Dell (2013). How NHTSA gets cars recalled. Available online at <https://www.edmunds.com/car-safety/how-nhtsa-gets-cars-recalled.html>.
- Kale, M. and A. Rastogi (2020, December). Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, Dublin, Ireland, pp. 97–102. Association for Computational Linguistics.
- Kleijnen, J. P. C. (2015). *Design and Analysis of Simulation Experiments* (2 ed.). New York: Springer.

- Kusner, M., Y. Sun, N. Kolkin, and K. Weinberger (2015). From word embeddings to document distances. In F. Bach and D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, pp. 957–966.
- Lareau, J. L. (2019). GM: We encourage employees, dealers to tattle after ignition switch crisis. Available online at <https://www.freep.com/story/money/cars/general-motors/2019/09/06/gm-ignition-switch-nhtsa-recalls-safety-defects/2099289001/>.
- Lavie, A. and A. Agarwal (2007). Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 228–231. Association for Computational Linguistics.
- Lee, T. Y. and E. T. BradLow (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research* 48(5), 881–894.
- Lim, K. W., W. Buntine, C. Chen, and L. Du (2016). Nonparametric Bayesian topic modelling with the hierarchical Pitman–Yor processes. *International Journal of Approximate Reasoning* 78, 172–191.
- Lin, C.-Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, Barcelona, Spain, pp. 74–81. Association for Computational Linguistics.
- Liu, A. X., Y. Liu, and T. Luo (2016). What drives a firm’s choice of product recall remedy? The impact of remedy cost, product hazard, and the CEO. *Journal of Marketing* 80(3), 79–95.
- Liu, J. and O. Toubia (2018). A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science* 37(6), 930–952.
- Liu, J., O. Toubia, and S. Hill (2021). Content-based model of web search behavior: An application to TV show search. *Management Science* 67(10), 6378–6398.
- Liu, X., P. V. Singh, and K. Srinivasan (2016). A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science* 35(3), 363–388.
- Liu, Y. and V. Shankar (2015). The dynamic impact of product-harm crises on brand preference and advertising effectiveness: An empirical analysis of the automobile industry. *Management Science* 61(10), 2514–2535.
- Mejia, J., S. Mankad, and A. Gopal (2021). Service quality using text mining: Measurement and consequences. *Manufacturing & Service Operations Management* 23(6), 1354–1372.

- Mukherjee, U. K., G. P. Ball, K. D. Wowak, K. V. Natarajan, and J. W. Miller (2022). Hiding in the herd: The product recall clustering phenomenon. *Manufacturing & Service Operations Management* 24(1), 392–410.
- Murray, S. (2022). Senator murray grills califf on FDA’s response to infant formula crisis, demands further action.
- Nam, H., Y. V. Joshi, and P. K. Kannan (2017). Harvesting brand information from social tags. *Journal of Marketing* 81(4), 88–108.
- Naumzik, C., S. Feuerriegel, and M. Weinmann (2022). I will survive: Predicting business failures from customer ratings. *Marketing Science* 41(1), 188–207.
- Netzer, O., R. Feldman, J. Goldenberg, and M. Fresko (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science* 31(3), 521–543.
- NHTSA (2015). NHTSA’s efforts to identify safety-related behicle defects. Available online at <https://www.oig.dot.gov/sites/default/files/NHTSA%20Safety-Related%20Vehicle%20Defects%20-%20Final%20Report%5E6-18-15.pdf>.
- NHTSA (2015). Office of inspector general audit report. Available online at <https://www.oig.dot.gov/sites/default/files/NHTSA%20Safety-Related%20Vehicle%20Defects%20-%20Final%20Report%5E6-18-15.pdf>.
- NHTSA (2019). Motor vehicle safety defects and recalls. Available online at https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/14218-mvsdefectsandrecalls_041619-v2-tag.pdf.
- Ni, J. and X. Huang (2018). Discovery-to-recall in the automotive industry: A problem-solving perspective on investigation of quality failures. *Journal of Supply Chain Management* 54(2), 71–95.
- Ni, Y., P. Müller, Y. Zhu, and Y. Ji (2018). Heterogeneous reciprocal graphical models. *Biometrics* 74(2), 606–615.
- Pennebaker, J. W., R. L. Boyd, K. Jordan, and K. Blackburn (2015). The development and psychometric properties of LIWC2015. Technical report.

- Phukan, R. (2017). Can cognitive predictive analytics prevent automobile recalls? Available online at <https://www.progress.com/blogs/can-cognitive-predictive-analytics-prevent-automobile-recalls>.
- Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog* 1(8), 9.
- Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21(140), 1–67.
- Rafieian, O. and H. Yoganarasimhan (2021). Targeting and privacy in mobile advertising. *Marketing Science* 40(2), 193–218.
- Raval, D. (2020). Whose voice do we hear in the marketplace? Evidence from consumer complaining behavior. *Marketing Science* 39(1), 168–187.
- Schmittlein, D. C., L. G. Cooper, and D. G. Morrison (1993). Truth in concentration in the land of (80/20) laws. *Marketing Science* 12(2), 167–183.
- Sezer, A. D. and Ç. Haksöz (2012). Optimal decision rules for product recalls. *Mathematics of Operations Research* 37(3), 399–418.
- Singh, K. and R. Grewal (2023). Lobbying and product recalls: A study of the u.s. automobile industry. *Journal of Marketing Research*, In Advance.
- Stout (2019). 2019 automotive defect & recall report. Available online at <https://www.stout.com/en/insights/report/2019-automotive-defect-and-recall-report>.
- Thirumalai, S. and K. K. Sinha (2011). Product recalls in the medical device industry: An empirical exploration of the sources and financial consequences. *Management Science* 57(2), 376–392.
- Tirunillai, S. and G. J. Tellis (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research* 51(4), 463–479.
- Toubia, O. (2020). A Poisson factorization topic model for the study of creative documents (and their summaries). *Journal of Marketing Research* 58(6), 1142–1158.
- Wallach, H. M., D. M. Mimno, and A. McCallum (2009). Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.), *Advances in*

Neural Information Processing Systems, NIPS 2009, Curran Associates, Rostrevor, Northern Ireland, pp. 1973–1981.

Wood, F. and Y. W. Teh (2009). A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation. In *Artificial Intelligence and Statistics*, pp. 607–614. Proceedings of Machine Learning Research.

Wowak, K. D., G. P. Ball, C. Post, and D. J. K. Jr. (2021). The influence of female directors on product recall decisions. *Manufacturing & Service Operations Management* 23(4), 895—913.

Zhang, M. and L. Luo (2023). Can consumer-posted photos serve as a leading indicator of restaurant survival? evidence from yelp. *Management Science* 69(1), 25–50.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi (2020, February). BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Zhang, X., Z. Qiao, L. Tang, P. W. Fan, E. A. Fox, and A. G. Wang (2016). Identifying product defects from user complaints: A probabilistic defect model. Report, Department of Computer Science, Virginia Polytechnic Institute & State.

Zhao, W., M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger (2019). MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 563–578. Association for Computational Linguistics.

Zhong, N. and D. A. Schweidel (2020). Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science* 39(4), 827–846.

Web Appendices

A The CRP Representation of the HDPYP

This section explains the underlying structure of the hierarchical CRP and introduces the notations related to the HDPYP. With the CRP metaphor (Buntine and Hutter, 2012), the PYP nodes in the HDPYP (i.e., μ_i , θ_i , and each element of $\{v_{i,1}, v_{i,2}, \dots, v_{i,J_i}\}$) are considered as *restaurants*, while all of the observed words in the textual data are treated as *customers*. In each restaurant, there are *tables*, and each customer is only allowed to sit at one table. Each table can only serve one *dish* (which indicates the topic in our setting), but the same dish can be served at multiple tables. Therefore, if we know both the table at which each customer is and the dish that is available at each table, then we know the dish (i.e., the topic) connected to each customer (i.e., the word).

As shown in Figure 1, for each recall i , the parent restaurants (i.e., the complaint nodes $v_{i,j}$ for $j \in \{1, 2, \dots, J_i\}$) generate the distribution of their child restaurant (i.e., the recall node θ_i). This means that each table at a child restaurant must be treated as a special customer who is sent back to one of the child restaurant's parent restaurants and is sitting at a table that serves the same dish as in the child restaurant. Technically, if a customer sits at a new table, a dish must be ordered from the global menu. In the CRP, ordering a dish from the global menu is equivalent to sending a new table to the corresponding parent restaurant (?). As a result, the parent restaurants have two types of customers (?): customers who arrived by themselves (denoted by *normal* customers) and customers who were sent from the child restaurant (denoted by *proxy* customers). In the HDPYP, the recall restaurants only have normal customers, whereas the complaint restaurants may have both types of customers simultaneously.

To express the model likelihood under the CRP, we need the following notations to characterize each restaurant's seating layout. For each recall i , we let $t_{i,k}^r$ denote the number of tables that serve the dish k , and $n_{i,k}^r$ denote the total number of normal customers eating dish k , for $k \in \{1, 2, \dots, K\}$. Then, $T_i^r = \sum_{k=1}^K t_{i,k}^r$ represents the total number of tables, and $N_i^r = \sum_{k=1}^K n_{i,k}^r$ represents the total number of normal customers across all K dishes of recall restaurant i . For the j th complaint of recall i , the corresponding $t_{i,j,k}^c$, $n_{i,j,k}^c$, $T_{i,j}^c$, and $N_{i,j}^c$ can be defined similarly. To capture the dependence of restaurant θ_i on its parent restaurants $v_{i,j}$ for $j \in \{1, 2, \dots, J_i\}$, we let $s_{i,j,k}^{r \rightarrow c}$ denote the number of tables serving dish k that are treated as proxy customers who are sent from recall restaurant θ_i to complaint restaurant $v_{i,j}$. Thus, $S_{i,j}^{r \rightarrow c} = \sum_{k=1}^K s_{i,j,k}^{r \rightarrow c}$ represents the total number of tables that are sent from recall restaurant i to complaint restaurant j across all K dishes. Because each table in the recall restaurant must be transformed into a

proxy customer and sent to the complaint restaurant, we have $T_i^r = \sum_{j=1}^{J_i} S_{i,j}^{r \rightarrow c} = \sum_{j=1}^{J_i} \sum_{k=1}^K S_{i,j,k}^{r \rightarrow c}$.

A.1 An Illustrative Example

For ease of understanding, we present a sample CRP seating layout of the HDPYP in Figure W1 that involves a child restaurant θ_i and its two parent restaurants $v_{i,j}$ and $v_{i,j'}$. In Figure W1, the proxy customers are transformed from their tables in the recall restaurant to tables in the parent restaurants (see the dashed-line arrows). We can see that there are three dishes (i.e., rice, corn, skewers) in the child restaurant. The number of tables associated with different dishes are $(t_{i,\text{rice}}^r, t_{i,\text{corn}}^r, t_{i,\text{skewers}}^r) = (2, 3, 2)$. The number of normal customers associated with different dishes are $(n_{i,\text{rice}}^r, n_{i,\text{corn}}^r, n_{i,\text{skewers}}^r) = (12, 13, 6)$. Thus, the total number of tables and customers are $T_i^r = 7$ and $N_i^r = 31$. The parent restaurant $v_{i,j}$ has four dishes (i.e., rice, corn, skewers, and tofu). Their numbers of tables and normal customers are

$$(t_{i,j,\text{rice}}^c, t_{i,j,\text{tofu}}^c, t_{i,j,\text{corn}}^c, t_{i,j,\text{skewers}}^c) = (1, 1, 1, 1),$$

$$(n_{i,j,\text{rice}}^c, n_{i,j,\text{tofu}}^c, n_{i,j,\text{corn}}^c, n_{i,j,\text{skewers}}^c) = (6, 4, 1, 3).$$

The parent restaurant $v_{i,j'}$ has three dishes (i.e., rice, corn, skewers). Its number of tables and of normal customers are

$$(t_{i,j',\text{rice}}^c, t_{i,j',\text{corn}}^c, t_{i,j',\text{skewers}}^c) = (1, 1, 2),$$

$$(n_{i,j',\text{rice}}^c, n_{i,j',\text{corn}}^c, n_{i,j',\text{skewers}}^c) = (7, 4, 7).$$

Note that the proxy customers, who are moved from their tables in the recall restaurant to tables in the complaint restaurants, have the same dish. In addition, the number of proxy customers who are sent to nodes j and j' are

$$(s_{i,j,\text{rice}}^{r \rightarrow c}, s_{i,j,\text{corn}}^{r \rightarrow c}, s_{i,j,\text{skewers}}^{r \rightarrow c}) = (1, 2, 1),$$

$$(s_{i,j',\text{rice}}^{r \rightarrow c}, s_{i,j',\text{corn}}^{r \rightarrow c}, s_{i,j',\text{skewers}}^{r \rightarrow c}) = (1, 1, 1).$$

Their summation is seven, corresponding to the number of tables in node θ_i .

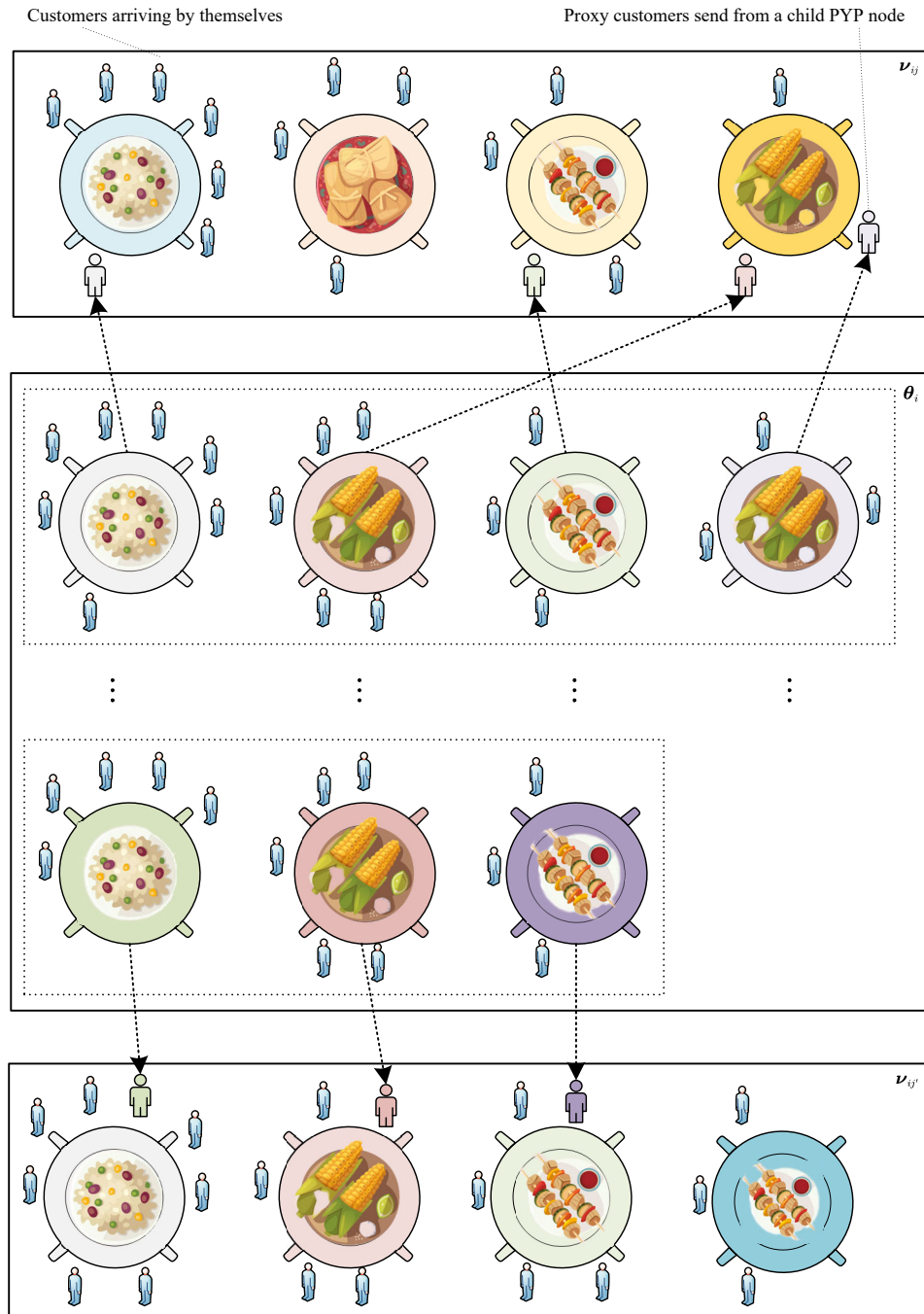


Figure W1: An Illustrative Example of the CRP Representation

A.2 Definitions of Table Indicators

The table indicator for a customer in the conventional CRP is a dummy random variable, which takes one when the customer contributes to creating a table in the associated restaurant and 0 otherwise (Chen et al., 2011). However, the table indicators in the HDPYP are much more complex due to the hierarchical dependence between the recall restaurants and the complaint restaurants. That is because if a customer contributes to the creation of a new table in the recall restaurant, he or she will also contribute to a particular complaint restaurant and thus assume the role of a proxy customer, sitting at one table in the complaint restaurant that serves the same dish.

We address this complication by introducing additional notations for the HDPYP. Let $u_{i,l}^r$ denote the table indicator of $w_{i,l}^r$, which represents the l th word in the i th recall statement. We define $u_{i,l}^r$ as a vector with two components, i.e., $u_{i,l}^r \triangleq (u_{i,l,1}^r, u_{i,l,2}^r)$. The first is an indicator variable that can take three possible values: $u_{i,l,1}^r = 0$, when $w_{i,l}^r$ does not contribute to the table count of the recall restaurant (and therefore also does not contribute to the table count of the complaint restaurant); $u_{i,l,1}^r = 1$, when $w_{i,l}^r$ only contributes to the table count of the recall restaurant; and $u_{i,l,1}^r = 2$, when $w_{i,l}^r$ contributes to opening tables in both the recall and the complaint restaurants. Accordingly, when $u_{i,l,1}^r \neq 0$, the second term $u_{i,l,2}^r$ indicates the complaint restaurant to which $w_{i,l}^r$ contributes a table.

Given the dish that each customer is eating, the probabilities of all of the possible scenarios can be computed as below:

$$p(u_{i,l,1}^r = 0 | z_{i,l}^r = k) = 1 - \frac{t_{i,k}^r}{n_{i,k}^r}, \quad (7)$$

$$p(u_{i,l,1}^r = 1, u_{i,l,2}^r = j | z_{i,l}^r = k) = \frac{t_{i,k}^r}{n_{i,k}^r} \left(1 - \frac{t_{i,j,k}^c}{n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c}} \right), \quad (8)$$

$$p(u_{i,l,1}^r = 2, u_{i,l,2}^r = j | z_{i,l}^r = k) = \frac{t_{i,k}^r}{n_{i,k}^r} \frac{t_{i,j,k}^c}{n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c}}, \quad (9)$$

where $t_{i,k}^r/n_{i,k}^r$ is the probability of customer $w_{i,l}^r$ opening a table in the recall restaurant i , and $t_{i,j,k}^c/(n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c})$ is the probability that the resulting proxy customer opens a table in the complaint restaurant j . Thus, the probability of customer $w_{i,l}^r$ opening a table in the recall restaurant with respect to dish k is the ratio of the number of tables serving dish k to the number of customers eating the dish.

Similarly, we let $u_{i,j,h}^c \in \{0, 1\}$ denote the table indicator of $w_{i,j,h}^c$, which represents the h th word in the j th complaint associated with the i th recall. Thus, $u_{i,j,h}^c = 0$, when $w_{i,j,h}^c$ does not contribute to the table count of the complaint restaurant; and $u_{i,j,h}^c = 1$, when $w_{i,j,h}^c$ does contribute to the table count of the complaint restaurant. Given the dish that each customer is eating, the probability distribution of $u_{i,j,h}^c$

can be derived as

$$p(u_{i,j,h}^c = 1 | z_{i,j,h}^c = k) = \frac{t_{i,j,k}^c}{n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c}}, \quad (10)$$

$$p(u_{i,j,h}^c = 0 | z_{i,j,h}^c = k) = 1 - \frac{t_{i,j,k}^c}{n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c}}. \quad (11)$$

That is because there are two types of customers who are eating dish k in a complaint restaurant: normal customers $n_{i,j,k}^c$, and proxy customers $s_{i,j,k}^{r \rightarrow c}$. Note that both types of customers can open new tables at the complaint restaurants.

B Inference Algorithm for the HDPYP

This appendix describes the inference algorithm for the HDPYP. Given the hyper-parameters and data, our goal is to estimate the following model parameters $\{\lambda, \mu, \nu, \theta, \Phi, \rho, z^r, z^c, u^r, u^c\}$.

B.1 Marginalized Model Likelihood

We first integrate out the following model parameters from the joint model likelihood, i.e., $\{\mu, \nu, \theta, \Phi\}$. The resulting joint conditional distribution can be derived as

$$\begin{aligned} & p(z^r, z^c, u^r, u^c, w^r | \alpha, a, b, \gamma, \delta, \sigma, \mathbf{x}, \lambda, \rho). \\ &= \prod_{k=1}^K \frac{\text{Beta}(M_k + \gamma)}{\text{Beta}(\gamma)} \prod_{i=1}^I \left[\prod_{j=1}^{J_i} \left(\frac{(b|a)_{T_{i,j}^c}}{(b)_{N_{i,j}^c + S_{i,j}^{r \rightarrow c}}} \prod_{k=1}^K S_{i,j,k}^{n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c}} \left(C_{i,j,k}^{n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c}} \right)^{-1} \right) \frac{\text{Beta}_K(\alpha + \sum_{j=1}^{J_i} t_{i,j}^c)}{\text{Beta}_K(\alpha)} \right] \\ & \prod_{i=1}^I \frac{\prod_{j=1}^{J_i} \Gamma(\exp(\lambda_i^\top x_{i,j}) + S_{i,j}^{r \rightarrow c})}{\Gamma(\sum_{j=1}^{J_i} (\exp(\lambda_i^\top x_{i,j}) + S_{i,j}^{r \rightarrow c}))} \frac{\Gamma(\sum_{j=1}^{J_i} \exp(\lambda_i^\top x_{i,j}))}{\prod_{j=1}^{J_i} \Gamma(\exp(\lambda_i^\top x_{i,j}))} \frac{(b|a)_{\sum_{j=1}^{J_i} S_{i,j}^{r \rightarrow c}}}{(b)_{N_i^r}} \left(\prod_{k=1}^K S_{\sum_{j=1}^{J_i} S_{i,j,k}^{r \rightarrow c}, a}^{n_{i,k}^r} \left(C_{\sum_{j=1}^{J_i} S_{i,j,k}^{r \rightarrow c}}^{n_{i,k}^r} \right)^{-1} \right) \\ & \prod_{i=1}^I \frac{\text{Beta}_K(\alpha + \sum_{j=1}^{J_i} s_{i,j}^{r \rightarrow c})}{\text{Beta}_K(\alpha)}, \end{aligned} \quad (12)$$

where $\text{Beta}_K(\cdot)$ ($\text{Beta}_{J_i}(\cdot)$) is a K (J_i) dimensional beta function that normalizes the Dirichlet distribution; $S_{A,a}^B$ is the generalized Stirling number given by the linear recursion (Buntine and Hutter, 2012); both $(b|a)_T$ and b_T denote the Pochhammer symbols (rising factorials) (?); M_k is a V -dimensional vector in which the v th component denotes the total number of words in the entire corpus that is the v th word in the dictionary and is assigned to topic k ; and $s_{i,j}^{r \rightarrow c} \triangleq (s_{i,j,1}^{r \rightarrow c}, s_{i,j,2}^{r \rightarrow c}, \dots, s_{i,j,K}^{r \rightarrow c})$ and $S_i^{r \rightarrow c} \triangleq (S_{i,1}^{r \rightarrow c}, S_{i,2}^{r \rightarrow c}, \dots, S_{i,J_i}^{r \rightarrow c})$ are respectively the K - and J_i -dimensional vectors, respectively, for the proxy customers sent from a recall restaurant to a complaint restaurant. The detailed derivation of Equation (12) is available from the authors upon request.

Because multiple customers can have the same dish at distinct tables and empty tables are not allowed at any restaurant (see Figure W1), the following constraints must be satisfied for a recall restaurant that has no child restaurant:

$$n_{i,k}^r \geq t_{i,k}^r > 0, \quad (13)$$

$$t_{i,k}^r = 0, \text{ if and only if } n_{i,k}^r = 0. \quad (14)$$

The first constraint implies that the number of customers having dish k must be greater than or equal to the number of tables serving that dish, where the equality holds when each table serving dish k has only one customer. The second constraint guarantees that an empty table is prohibited. Given that the complaint restaurants might have two types of customers, the constraints for a given complaint restaurant are as follows:

$$n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c} \geq t_{i,j,k}^c > 0, \quad (15)$$

$$t_{i,j,k}^c = 0, \text{ if and only if } n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c} = 0, \quad (16)$$

which incorporate the proxy customers $s_{i,j,k}^{r \rightarrow c}$ compared with their counterparts in Equations (13)-(14).

B.2 Removing the Current Topic Assignment

Before sampling a new topic indicator for a word, we need to remove its current value from the related statistics according to its table indicator. However, the table indicator for any word is not recorded. Therefore, table indicators need to be randomly assigned by sampling. Because there are two types of words (i.e., normal and proxy), we summarize their inference procedures separately in Algorithms 1 and 2.

Algorithm 1 describes how the HDPYP samples table indicator $u_{i,l}^r$ for word $w_{i,l}^r$ in the recall statements, along with how relevant count statistics change when removing its value. We have proved in Online Appendix B.5 that all of the constraints are automatically satisfied in Algorithm 1. For recall restaurants, removing a customer, say $w_{i,l}^r$, with current dish $z_{i,l}^r = k$, amounts to decreasing the number of customers $n_{i,k}^r$ by one and thus the associated customer count. The table indicator $u_{i,l}^r = (u_{i,l,1}^r, u_{i,l,2}^r)$ of $w_{i,l}^r$ is randomly sampled according to the probability each time that $w_{i,l}^r$ is removed over the course of the Gibbs sampling. Given the definition of the table indicators introduced above, there are only $3J_i$ possible values of $u_{i,l}^r$. We obtain the probability of each value using Equations (7)-(9), and sample $u_{i,l}^r$

accordingly. If the sampled $u_{i,l}^r = (0, j)$, $w_{i,l}^r$ does not contribute to opening a table at any restaurant, the table count remains the same. If the sampled $u_{i,l}^r = (1, j)$, $w_{i,l}^r$ contributes to opening a table but only for the recall restaurant, both the number of tables $t_{i,k}^r$ in the recall restaurants and the number of proxy customers $s_{i,j,k}^{r \rightarrow c}$ decrease by one. Finally, if the sampled $u_{i,l}^r = (2, j)$, we decrement the table counts of both the recall and the complaint restaurants by one.

Algorithm 1: Sample to remove the current topic assignment for word $w_{i,l}^r$

```

1 for  $j = 1$  to  $J_i$  complaint, given  $z_{i,l}^r = k$  do
2   if  $n_{i,k}^r > t_{i,k}^r = 1$  then
3      $p(u_1^r = 0, u_2^r = j) = 1$ ;
4   else
5     compute  $\{p(u_1^r = 0, u_2^r = j), p(u_1^r = 1, u_2^r = j), p(u_1^r = 2, u_2^r = j)\}$  by (7), (8) and (9);
6   end
7 end
8 Sample  $u_{i,l}^r = (u_1^r, u_2^r)$  by computed  $\{p(u_1^r = 0, u_2^r = j), p(u_1^r = 1, u_2^r = j), p(u_1^r = 2, u_2^r = j)\}_{j=1}^{J_i}$ 
   via Discrete distribution;
9 if  $u_1^r = 0$  given  $u_2^r = j$  then
10  decrement  $n_{i,k}^r$ ;
11 else if  $u_1^r = 1$  then
12  decrement  $n_{i,k}^r, t_{i,k}^r$  and  $s_{i,j,k}^{r \rightarrow c}$ ;
13 else if  $u_1^r = 2$  then
14  decrement  $n_{i,k}^r, t_{i,k}^r, s_{i,j,k}^{r \rightarrow c}$  and  $t_{i,j,k}^c$ ;
15 end
16 Decrement  $M_{i,k,w_{i,l}^r}$  and update other related statistics including  $N_i^r, T_i^r, S_{i,j}^{r \rightarrow c}, T_{i,j}^c, M_{k,w_{i,l}^r}$ .

```

Algorithm 2 describes the procedure for removing a complaint word $w_{i,j,l}^c$. We have proved in Online Appendix B.5 that all of the constraints are automatically satisfied under Algorithm 2. The process of removing a word $w_{i,j,h}^c$ from the corresponding complaint restaurant j is comparatively easy, as we only need to know whether the word is responsible for opening a table for the restaurant, without regard to its impact on the parent restaurant. In the same vein, as we obtain the value of table indicator $u_{i,l}^r$ for $w_{i,l}^r$ in the recall restaurants, we sample $u_{i,j,h}^c \in \{0, 1\}$ according to the probabilities computed in Equations (10) and (11).

B.3 Sampling a New Topic Assignment

After decrementing the topic associated with a word, we use a blocked Gibbs sampler to sample a new topic for the word and update the corresponding customer counts and table counts. The conditional posterior used in the sampling can be computed quickly when the full posterior is represented in a modularized form.

We first elaborate on the words in the recall statements. Suppose that a topic k^* is sampled for a

Algorithm 2: Sample to remove the current topic assignment for word $w_{i,j,h}^c$

- 1 **if** $n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c} > t_{i,j,k}^c = 1$ **then**
 - 2 | $p(u_{i,j,h}^c = 0) = 1$;
 - 3 **else**
 - 4 | compute $\{p(u_{i,j,h}^c = 0), p(u_{i,j,h}^c = 1)\}$ given $z_{i,j,h}^c = k$ by (10) and (11);
 - 5 **end**
 - 6 Sample $u_{i,j,h}^c$ by computed $\{p(u_{i,j,h}^c = 0), p(u_{i,j,h}^c = 1)\}$ via Discrete distribution;
 - 7 **if** $u_{i,j,h}^c = 0$ **then**
 - 8 | decrement $n_{i,j,k}^c$;
 - 9 **else if** $u_{i,j,h}^c = 1$ **then**
 - 10 | decrement $n_{i,j,k}^c$ and $t_{i,j,k}^c$;
 - 11 **end**
 - 12 Decrement $M_{i,k,w_{i,j,h}^c}$ and update other related statistics including $N_{i,j}^c$, $T_{i,j}^c$ and $M_{k,w_{i,j,h}^c}$.
-

word $w_{i,l}^r$ in the i th recall statement, so $z_{i,l}^r = k^*$. Given the possible values that the table indicator $u_{i,l}^r = (u_{i,l,1}^r, u_{i,l,2}^r)$ of word $w_{i,l}^r$ can take, we derive the joint conditional probabilities of the topic assignment and its corresponding table indicator as follows:

$$p\left(z_{i,l}^r = k^*, u_{i,l,1}^r = 0, u_{i,l,2}^r = j \mid z_{i,l}^r = z_{i,l}^r, z^c, w^r, w^c, u^{r \rightarrow c}, u^c, \alpha, \gamma, a, b, \delta, \sigma, \mathbf{x}\right) \\ \propto \frac{M_{k^*, w_{i,l}^r} + \gamma_{w_{i,l}^r}}{\sum_{v=1}^V (M_{k^*, v} + \gamma_v)} \frac{1}{b + N_i^r} \frac{S_{\sum_{j'=1}^{J_i} s_{i,j',k^*}^{r \rightarrow c}, a}^{n_{i,k^*}^r + 1}}{S_{\sum_{j'=1}^{J_i} s_{i,j',k^*}^{r \rightarrow c}, a}^{n_{i,k^*}^r}} \frac{n_{i,k^*}^r + 1 - \sum_{j'=1}^{J_i} s_{i,j',k^*}^{r \rightarrow c}}{n_{i,k^*}^r + 1}, \quad (17)$$

$$p\left(z_{i,l}^r = k^*, u_{i,l,1}^r = 1, u_{i,l,2}^r = j \mid z_{i,l}^r = z_{i,l}^r, z^c, w^r, w^c, u^{r \rightarrow c}, u^c, \alpha, \gamma, a, b, \delta, \sigma, \mathbf{x}\right) \\ \propto \frac{M_{k^*, w_{i,l}^r} + \gamma_{w_{i,l}^r}}{\sum_{v=1}^V (M_{k^*, v} + \gamma_v)} \frac{1}{b + N_{i,j}^c + S_{i,j}^{r \rightarrow c}} \frac{S_{t_{i,j,k^*}^c + a}^{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c} + 1}}{S_{t_{i,j,k^*}^c + a}^{n_{i,j,k^*}^c}} \frac{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c} + 1 - t_{i,j,k^*}^c}{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c} + 1} \frac{\exp(\lambda_i^T \mathbf{x}_{i,j}) + S_{i,j}^{r \rightarrow c, -z_{i,l}^r}}{\sum_{j'=1}^{J_i} (\exp(\lambda_i^T \mathbf{x}_{i,j'}) + S_{i,j'}^{r \rightarrow c, -z_{i,l}^r})} \\ \frac{b + a \sum_{j'=1}^{J_i} S_{i,j'}^{r \rightarrow c}}{b + N_i^r} \frac{S_{\sum_{j'=1}^{J_i} s_{i,j',k^*}^{r \rightarrow c}, a}^{n_{i,k^*}^r + 1}}{S_{\sum_{j'=1}^{J_i} s_{i,j',k^*}^{r \rightarrow c}, a}^{n_{i,k^*}^r}} \frac{\sum_{j'=1}^{J_i} s_{i,j',k^*}^{r \rightarrow c} + 1}{n_{i,k^*}^r + 1} \frac{\alpha_{k^*} + \sum_{j'=1}^{J_i} s_{i,j',k^*}^{r \rightarrow c}}{\sum_{k=1}^K (\alpha_k + \sum_{j'=1}^{J_i} s_{i,j',k}^{r \rightarrow c})}, \quad (18)$$

$$p\left(z_{i,l}^r = k^*, u_{i,l,1}^r = 2, u_{i,l,2}^r = j \mid z_{i,l}^r = z_{i,l}^r, z^c, w^r, w^c, u^{r \rightarrow c}, u^c, \alpha, \gamma, a, b, \delta, \sigma, \mathbf{x}\right) \\ \propto \frac{M_{k^*, w_{i,l}^r} + \gamma_{w_{i,l}^r}}{\sum_{v=1}^V (M_{k^*, v} + \gamma_v)} \frac{b + a T_{i,j}^c}{b + N_{i,j}^c + S_{i,j}^{r \rightarrow c}} \frac{S_{t_{i,j,k^*}^c + a}^{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c} + 1}}{S_{t_{i,j,k^*}^c + a}^{n_{i,j,k^*}^c}} \frac{t_{i,j,k^*}^c + 1}{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c} + 1} \frac{\alpha_{k^*} + \sum_{j=1}^{J_i} t_{i,j,k^*}^c}{\sum_{k=1}^K (\alpha_k + \sum_{j=1}^{J_i} t_{i,j,k^*}^c)} \\ \frac{\exp(\lambda_i^T \mathbf{x}_{i,j}) + S_{i,j}^{r \rightarrow c}}{\sum_{j=1}^{J_i} (\exp(\lambda_i^T \mathbf{x}_{i,j}) + S_{i,j}^{r \rightarrow c})} \frac{b + a \sum_{j=1}^{J_i} S_{i,j}^{r \rightarrow c}}{b + N_i^r} \frac{S_{\sum_{j=1}^{J_i} s_{i,j,k^*}^{r \rightarrow c}, a}^{n_{i,k^*}^r + 1}}{S_{\sum_{j=1}^{J_i} s_{i,j,k^*}^{r \rightarrow c}, a}^{n_{i,k^*}^r}} \frac{\sum_{j=1}^{J_i} s_{i,j,k^*}^{r \rightarrow c} + 1}{n_{i,k^*}^r + 1} \frac{\alpha_{k^*} + \sum_{j=1}^{J_i} s_{i,j,k^*}^{r \rightarrow c}}{\sum_{k=1}^K (\alpha_k + \sum_{j=1}^{J_i} s_{i,j,k^*}^{r \rightarrow c})}. \quad (19)$$

Note that we need to distinguish the following two cases when sampling the table indicator for $w_{i,l}^r$:

- Case 1: $n_{i,k^*}^r = 0$. This case implies that dish k^* is a new dish and $t_{i,k^*}^r = 0$ due to constraint (14). Thus, assigning customer $w_{i,l}^r$ to eat dish k^* first creates a new table serving dish k^* in the recall restaurant, leading us to rule out the situation $u_l^r = 0$ immediately. However, whether the new table triggers the creation of a new table in the j th complaint restaurant depends on the exact value of the sampled table indicator of $u_{i,l}^r$. Accordingly, we determine $u_{i,l}^r$ using Equations (18) and (19).
- Case 2: $n_{i,k^*}^r > 0$. In contrast, this case does not imply that dish k^* is new. Thus, we determine $u_{i,l}^r$ in light of Equations (17), (18), and (19) together.

We now describe the procedure for the words in consumer complaints. Similarly, for each word $w_{i,j,h}^c$, we derive the following joint conditional probabilities of a topic assignment $z_{i,j,h}^c = k^*$ and its associated table indicator $u_{i,j,h}^c$:

$$p\left(z_{i,j,h}^c = k^*, u_{i,j,h}^c = 0 \mid z^r, z^{c,-z_{i,j,h}^c}, w^r, w^c, u^r, u^{c,-u_{i,j,h}^c}, \alpha, \gamma, a, b, \lambda\right) \\ \propto \frac{M_{k^*, w_{i,j,h}^c} + \gamma w_{i,j,h}^c}{\sum_{v=1}^V (M_{k^*, v} + \gamma v)} \frac{1}{b + N_{i,j}^c + S_{i,j}^{r \rightarrow c}} \frac{S_{i,j,k^*, a}^{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c} + 1}}{S_{i,j,k^*, a}^{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c}}} \frac{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c} + 1 - t_{i,j,k^*}^c}{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c} + 1}, \quad (20)$$

$$p\left(z_{i,j,h}^c = k^*, u_{i,j,h}^c = 1 \mid z^r, z^{c,-z_{i,j,h}^c}, w^r, w^c, u^r, u^{c,-u_{i,j,h}^c}, \alpha, \gamma, a, b, \lambda\right) \\ \propto \frac{M_{k^*, w_{i,j,h}^c} + \gamma w_{i,j,h}^c}{\sum_{v=1}^V (M_{k^*, v} + \gamma v)} \frac{b + a T_{i,j}^c}{b + N_{i,j}^c + S_{i,j}^{r \rightarrow c}} \frac{S_{i,j,k^*, a}^{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c} + 1}}{S_{i,j,k^*, a}^{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c}}} \frac{t_{i,j,k^*}^c + 1}{n_{i,j,k^*}^c + s_{i,j,k^*}^{r \rightarrow c} + 1} \frac{\alpha_{k^*} + \sum_{j=1}^J t_{i,j,k^*}^c}{\sum_{k=1}^K (\alpha_k + \sum_{j=1}^J t_{i,j,k}^c)}. \quad (21)$$

There are also two cases for the words in consumer complaints. When $n_{i,j,k^*} = 0$, it means that the corresponding complaint restaurant has no a table serving dish k^* , and therefore the word $w_{i,j,h}^c$ must open a new table with dish k^* , so implicitly, $w_{i,j,h}^c = 1$. When $n_{i,j,k^*} = 1$, we determine its table contribution by the sampled values of the table indicator according to Equations (20) and (21).

B.4 Approximate Estimation of the Other Model Parameters

Using the sampled u and z , we can reconstruct the PYPs from the associated customer counts and table counts. Recovering the full posterior distribution of the PYPs is a complicated task. Thus, we use the expected values of their conditional marginal posterior distributions (i.e., the Dirichlet distribution or the PYP) via sampling, as in Lim et al. (2016). Note that the parameter λ is estimated by optimizing the full log-likelihood of the Dirichlet-multinomial regression model in Equation (5) given the estimated ρ and the data. The expressions of the other model parameters are shown below.

$$\widehat{\boldsymbol{\mu}}_{i,k} = \mathbb{E}_{z_i^c, t_i^c | w_i^c, \alpha, a, b, \gamma, \delta, \sigma, \mathbf{x}_i} \left[\frac{\alpha_k + \sum_{j=1}^{J_i} t_{i,j,k}^c}{\sum_{k=1}^K \left(\alpha_k + \sum_{j=1}^{J_i} t_{i,j,k}^c \right)} \right]. \quad (22)$$

$$\widehat{\mathbf{v}}_{i,j,k} = \mathbb{E}_{z_{i,j}^c, t_{i,j}^c, s_{i,j}^{r \rightarrow c} | w_{i,j}^c, \alpha, a, b, \gamma, \delta, \sigma, \mathbf{x}_{i,j}} \left[\frac{\left(n_{i,j,k} + s_{i,j,k}^{r \rightarrow c} \right) - a \times t_{i,j,k}^c}{b + N_{i,j}^c + S_{i,j}^{r \rightarrow c}} + \mu_{i,k} \frac{a \times T_{i,j}^c + b}{b + N_{i,j}^c + S_{i,j}^{r \rightarrow c}} \right]. \quad (23)$$

$$\widehat{\boldsymbol{\rho}}_{i,j} = \mathbb{E}_{z_i^r, s_{i,j}^{r \rightarrow c} | w_i^r, \alpha, a, b, \gamma, \delta, \sigma, \mathbf{x}_{i,j}} \left[\frac{\exp \left(\lambda_i^T \mathbf{x}_{i,j} \right) + S_{i,j}^{r \rightarrow c}}{\sum_{j=1}^{J_i} \left(\exp \left(\lambda_i^T \mathbf{x}_{i,j} \right) + S_{i,j}^{r \rightarrow c} \right)} \right]. \quad (24)$$

$$\widehat{\boldsymbol{\theta}}_{i,k} = \mathbb{E}_{z_i^r, z_i^c, t_i^r, t_i^c | w_i^r, w_i^c, \alpha, a, b, \gamma, \delta, \sigma, \mathbf{x}_i} \left[\frac{\left(a \times T_i^r + b \right) \left(\sum_{j=1}^{J_i} \widehat{\boldsymbol{\rho}}_{i,j} v_{i,j,k} \right) + n_{i,k}^r - \left(a \times t_{i,k}^r \right)}{b + N_i^r} \right]. \quad (25)$$

$$\widehat{\boldsymbol{\phi}}_{k,v} = \mathbb{E}_{z_{1:l}^r, z_{1:l}^c, t_{1:l}^r, t_{1:l}^c | w_{1:l}^r, w_{1:l}^c, \alpha, a, b, \gamma, \delta, \sigma, \mathbf{x}_{1:l}, J_i} \left[\frac{\gamma_v + M_{k,v}}{\sum_{v=1}^V \left(\gamma_v + M_{k,v} \right)} \right]. \quad (26)$$

B.5 Properties and Proofs

Proposition 1. *When removing a recall word via Algorithm 1, the conditions (13)-(16) for the recall and complaint restaurants are always satisfied.*

Proof. We first focus on the impact of removing word $w_{i,l}^r$ on the recall restaurant, followed by the impact of doing so on the complaint restaurant. For the recall restaurant, we identify the following situations of recall restaurant before removing $w_{i,l}^r$, assuming the topic of $w_{i,l}^r, z_{i,l}^r = k$:

1. $n_{i,k}^r = t_{i,k}^r > 0$, which means that each table in the i th recall restaurant has only one customer. In this situation, the first element of the table indicator $u_{i,l}^r, u_{i,l,1}^r$, which corresponds to the contribution of opening a table associated with $w_{i,l}^r$, can be either 1 or 2. For the case $u_{i,l,1}^r = 1$, which implies that $w_{i,l}^r$ only contributes to opening a table for the recall restaurant and not for the complaint restaurant, we can infer the current state of complaint restaurant $n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c} > t_{i,j,k}^c$ (note that if $n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c} = t_{i,j,k}^c$, $u_{i,l,1}^r$ should be 2, even though $u_{i,l,1}^r$ is already 1). In this case, it is easy to see that (13)-(16) still hold when removing $w_{i,l}^r$, because count $n_{i,k}^r, t_{i,k}^r$ and $s_{i,j,k}^{r \rightarrow c}$ decrements by 1 simultaneously via Algorithm 1; see the twelfth line of the algorithm. For the case $u_{i,l,1}^r = 2$, which suggests that $w_{i,l}^r$ contributes to creating a table for both the recall and the complaint restaurants, we have the state of complaint restaurant $n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c} \geq t_{i,j,k}^c$ before removing $w_{i,l}^r$. In this case, it is not difficult to see that (13)-(16) are also sufficient because Algorithm 1 decrements the count $n_{i,k}^r, t_{i,k}^r, s_{i,j,k}^{r \rightarrow c}$ and $t_{i,j,k}^c$; see the fourteenth line of the algorithm.

2. $n_{i,k}^r > t_{i,k}^r > 1$, in which case a customer must be removed from $n_{i,k}^r$, but a table can either be

removed or not, depending on the value of the sampled table indicator. Following the logic of the analysis in situation 1, it is not difficult to see that the constraints (13)-(16) remain satisfied.

3. $n_{i,k}^r > t_{i,k}^r = 1$, in which case there is only one table exists in the recall restaurant and more than one customer sitting at it. In this situation, the table must not be removed, or constraint (14) is not satisfied because there are other customers sitting at that table and sharing the dish with $w_{i,l}^r$. To circumvent this problem, Algorithm 1 mandates $p(u_1^r = 0, u_2^r = j) = 1$; see the third line of the algorithm. \square

Proposition 2. *When removing a complaint word via Algorithm 2, conditions (13)-(16) for the recall and complaint restaurants are always satisfied.*

Proof. We first point out that constraints (13)-(14) on the recall restaurant always hold, as removing the complaint word $w_{i,j,l}^c$ has no impact on the recall restaurant. Next, we show that constraints (15)-(16) on the complaint restaurant hold when implementing Algorithm 2. We identify the following three situations before removing $w_{i,j,l}^c$:

1. $n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c} = t_{i,j,k}^c > 0$, in which case each table in the complaint restaurant has only one customer, i.e., each customer sits at a separate table. In this situation, the table indicator of word $w_{i,j,l}^c, u_{i,j,l}^c$, corresponding to the contribution of opening a table at a complaint restaurant can only be 1. In other words, $w_{i,j,l}^c$ opens a table. We see that the constraints (15)-(16) still suffice when removing $w_{i,j,l}^c$, because Algorithm 2 decrements $n_{i,j,k}^c$ and $t_{i,j,k}^c$ simultaneously; see the tenth line of the algorithm.

2. $n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c} > t_{i,j,k}^c > 1$, in which case $w_{i,j,l}^c$ has a chance of opening a table in a complaint restaurant. In this situation, the table indicator $u_{i,j,l}^c$ can be either 0 or 1. In the case $u_{i,j,l}^c = 0$, suggesting that $w_{i,j,l}^c$ does not open a table, Algorithm 2 only decrements customer count $n_{i,j,k}^c$ when removing $w_{i,j,l}^c$ (see the eighth line of the algorithm), so (15)-(16) are not violated. In the case $u_{i,j,l}^c = 1$, Algorithm 2 decrements both customer count $n_{i,j,k}^c$ and $t_{i,j,k}^c$ when removing $w_{i,j,l}^c$ (see the tenth line of the algorithm), we see that (15)-(16) remain satisfied.

3. $n_{i,j,k}^c + s_{i,j,k}^{r \rightarrow c} > t_{i,j,k}^c = 1$, corresponding to the situation in which many customers sit at the only table. In this situation, the table cannot be removed. Algorithm 2 guards against this problem by mandating that no customer can create the table; see the second line of the algorithm. \square

C The Stochastic Kriging Metamodeling Method

Metamodels are mathematical approximations of the input/output (I/O) relationship implied by the underlying simulation model (Kleijnen, 2015), and they can be used to efficiently predict the unexplored parameter space. We let $\Xi \in \Omega$ denote the hyper-parameter combination, and $\Omega \subseteq \mathbb{R}^{|\Xi|}$ represent the

experimental region. Without loss of generality, let $\mathcal{Y}(\Xi)$ denote the goodness-of-fit measure of interest. With these notations, the hyper-parameter optimization problem can be expressed as

$$\Xi^* = \arg \min_{\Xi \in \Omega} \mathbb{E}[\mathcal{Y}(\Xi)], \quad (27)$$

where $\mathbb{E}[\cdot]$ implies that we care about the expected value of $\mathcal{Y}(\Xi)$ due to sampling uncertainty.

The hyper-parameters in the HDPYP include $\Xi = (a, b, \alpha, \gamma, \sigma, K)$. We use *perplexity* as our target metric for optimization because it is most commonly used to evaluate the goodness-of-fit of a topic model. The perplexity for the HDPYP can be computed as

$$\exp \left(- \frac{\sum_{i=1}^I \left(\sum_{l=1}^{L_i} \log p \left(w_{i,l}^r | \theta_i, \Phi \right) + \sum_{j=1}^{J_i} \sum_{h=1}^{H_{i,j}} \log p \left(w_{i,j,h}^c | \mathbf{v}_{i,j}, \Phi \right) \right)}{\sum_{i=1}^I \left(L_i + \sum_{j=1}^{J_i} H_{i,j} \right)} \right), \quad (28)$$

where $p(w_{i,l}^r | \theta_i, \Phi) = \sum_{k=1}^K \phi_{k,w_{i,l}} \theta_{i,k}$ and $p(w_{i,j,h}^c | \mathbf{v}_{i,j}, \Phi) = \sum_{k=1}^K \phi_{k,w_{i,j,h}} \mathbf{v}_{i,j,k}$ are the likelihood of sampling the words $w_{i,l}$ and $w_{i,j,h}$ given the associated document-topic distribution and topic-word distributions.

Following Ankenman et al. (2010), the output of $\mathcal{Y}(\Xi)$ after an algorithmic execution on replication d ($d = 1, 2, \dots, D$) at parameter combination Ξ is modeled as

$$\mathcal{Y}_d(\Xi) = Y(\Xi) + \varepsilon_d(\Xi). \quad (29)$$

The second term $\varepsilon_d(\Xi)$ denotes the random noise of the simulation, which is used to characterize the sampling uncertainties in the HDPYP and is commonly assumed to be i.i.d across replications with mean zero. The first term $Y(\Xi)$ models the true part of the perplexity output at Ξ , which is specified as

$$Y(\mathbf{x}) = \mathbf{f}(\Xi)^\top \boldsymbol{\beta} + M(\Xi), \quad (30)$$

where $\mathbf{f}(\cdot)$ is a $p \times 1$ vector of known specified functions, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $M(\cdot)$ is the so-called Gaussian random field with mean zero (?). Note that $M(\cdot)$ exhibits spatial correlation, implying that the values of $M(\Xi)$ and $M(\Xi')$ are similar if Ξ and Ξ' are close to each other. The spatial correlation is measured by the covariance function $\text{Cov}[M(\Xi), M(\Xi')] \triangleq \Sigma_M(\Xi, \Xi')$, which is a variance-covariance matrix.

Suppose that the HDPYP model has been run at Q distinct parameter combinations $\Xi_1, \Xi_2, \dots, \Xi_Q$ D times, which yield the $Q \times 1$ vector of the observed perplexity output $\bar{\mathcal{Y}} = (\bar{\mathcal{Y}}(\Xi_1), \bar{\mathcal{Y}}(\Xi_2), \dots, \bar{\mathcal{Y}}(\Xi_Q))^\top$,

where the q th element $\bar{\mathcal{Y}}(\Xi_q) = (1/D) \sum_{d=1}^D \mathcal{Y}_d(\Xi_q)$, for $q = 1, 2, \dots, Q$. We are now interested in predicting the perplexity output at a new parameter combination $\Xi_0 \in \Omega$. The SK prediction at Ξ_0 is given by

$$\hat{Y}(\Xi_0) = \mathbf{f}(\Xi_0)^\top \boldsymbol{\beta} + \boldsymbol{\Sigma}_M(\Xi_0, \Xi)^\top (\boldsymbol{\Sigma}_M(\Xi, \Xi') + \boldsymbol{\Sigma}_\epsilon)^{-1} (\bar{\mathcal{Y}} - \mathbf{F}\boldsymbol{\beta}), \quad (31)$$

where $\boldsymbol{\Sigma}_M(\Xi_0, \Xi)$ is the $K \times 1$ vector of the spatial covariances between Ξ and the predicted combination Ξ_0 ; $\boldsymbol{\Sigma}_\epsilon$ denotes the $K \times K$ covariance matrix with element $\text{Cov}[\sum_{d=1}^D \epsilon_d(\Xi_k)/D, \sum_{d=1}^D \epsilon_d(\Xi_{k'})/D]$; and $\mathbf{F} = [\mathbf{f}(\Xi_1), \mathbf{f}(\Xi_2), \dots, \mathbf{f}(\Xi_Q)]^\top$. In practice, the parameters $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_M$ and $\boldsymbol{\Sigma}_\epsilon$ must be estimated from the perplexity outputs obtained. Because the values of the parameter combinations are taken in a certain region, we can easily predict the perplexity for any combinations that have not been explored in the experiment. Interested readers are directed to refer to [Ankenman et al. \(2010\)](#) and ? for the technical details.

D Optimal Hyper-Parameters

To apply metamodel-based methods such as the SK method, it is important to know the experimental range (i.e., the low and high levels) of each parameter to quantify its impact on the performance metric (i.e., perplexity). In line with the related studies (see e.g., ??), we determine the respective low and high values of $(\alpha, \gamma, \sigma^2, a, b, K)$ for the HDPYP. In other words, $\alpha \in [0.01, 0.20]$, $\gamma \in [0.005, 0.020]$, $\sigma^2 \in [0.00, 1.00]$, $a \in [0.00, 0.50]$, $b \in [0.00, 20.00]$, and $K \in [5, 25]$. We further transform the parameters' low and high levels to two standardized values, namely, -1 and 1 , to remove the potential impact of their original units and ranges on the optimization.

We adopt *Latin hypercube sampling* (LHS) to experiment with 60 combinations of the six parameters. LHS is a type of stratified sampling technique inherited from traditional *Latin square designs* ([Kleijnen, 2015](#), pp 199-203). A well-known rule-of-thumb number of combinations for LHS in the SK method is 10 times the number of parameters (?). More specifically, we begin by splitting the range of each parameter into 60 ($= 10 \times 6$) mutually exclusive and exhaustive intervals of equal probability. Later on, we randomly sample one value for a parameter from each interval, without replacement, which yields 60 values of each parameter. We round the sampling values of K because it can only take integers. Finally, we pair the 60 values of one parameter with the 60 values of the other parameters one by one, randomly, and without replacement until a set of 60 experimental combinations is formed. We run five replications at each of the 60 experimental combinations to reduce sampling uncertainty. [Tables W1](#) and [W2](#) report the 60 experimental combinations, along with its resulting average perplexity and its standard

deviation across the five replications.

Upon obtaining the experimental results, we fit an SK model. Leveraging the fitted model, we estimate the perplexity scores over a wide unexplored space of $(\alpha, \gamma, \sigma^2, a, b, K)$. The minimum predicted value of the perplexity score is determined by the SK method at parameter combination (0.20, 0.019, 0.27, 0.02, 2.10, 23), with the optimal predicted perplexity $\hat{\gamma}^* = 597.87$. Because the optimal parameter setting is determined by applying the SK method, it is necessary to confirm the optimality through multiple replications of the HDPYP model under the optimal setting. We compare the simulation outputs between the minimum perplexities obtained at one of the 60 experimental combinations in Table W1 (i.e., Comb. 20) and those obtained at the optimal combination that is predicted. Figure W2 displays the corresponding outputs through five independent replications. We find that the five optima with a mean perplexity value of 599.04 are at the desired levels, as they are very close to the predicted $\hat{\gamma}^* = 597.87$ and below the smallest 603.46 of the 60 experimental settings.

We apply a similar procedure to determine the optimal hyper-parameters of the LDA used in our study as a benchmark topic model. For the LDA, we treat all of the consumer complaints and recall statements as separate documents. We find that 23 is also the optimal number of topics for the LDA, and the labels of these 23 topics are similar to these 23 topics extracted by the HDPYP.

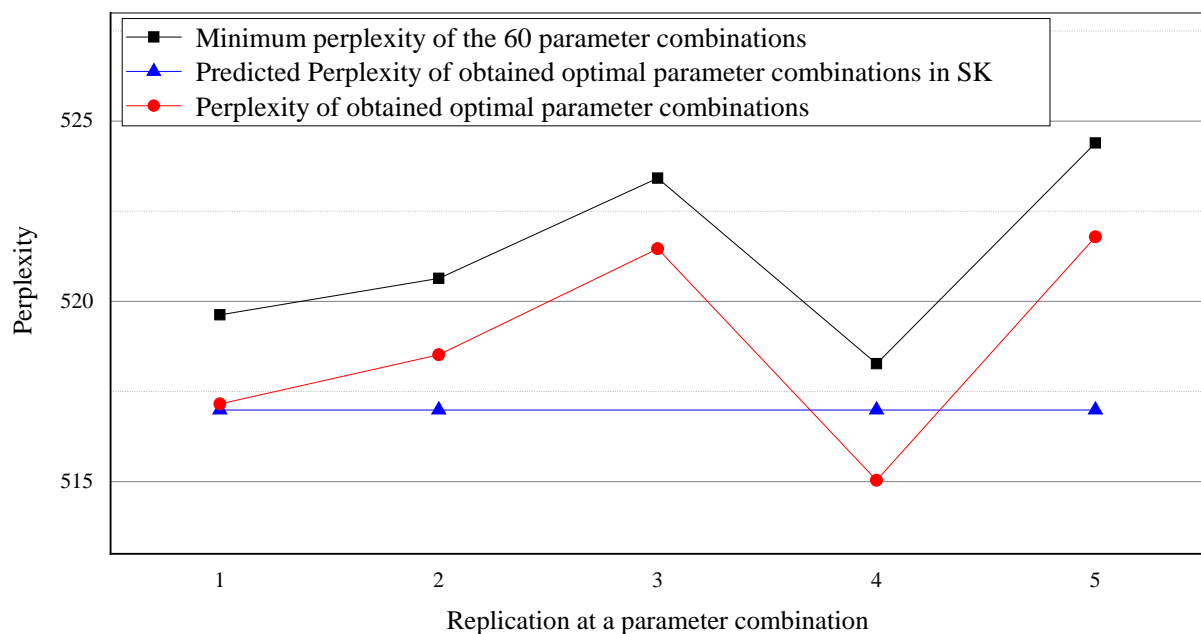


Figure W2: Comparison of the Perplexity between Experimental Setting and the Optimal Setting

Table W1: Experimental Parameter Combinations and Resulting Perplexity

Comb.	α	γ	σ^2	a	b	K	$\bar{\mathcal{Y}}$	std(\mathcal{Y})
1	0.11	0.013	0.62	0.09	19.15	10	730.52	3.45
2	0.20	0.008	0.78	0.20	18.07	6	787.47	3.04
3	0.02	0.012	0.25	0.33	2.97	10	687.41	3.88
4	0.16	0.016	0.11	0.08	2.25	17	626.61	2.62
5	0.15	0.006	0.91	0.03	19.96	19	666.66	1.01
6	0.11	0.013	0.94	0.49	16.76	11	738.15	1.35
7	0.19	0.012	0.69	0.31	11.56	24	628.22	0.95
8	0.04	0.006	0.82	0.44	5.55	23	621.17	0.69
9	0.05	0.016	0.68	0.13	8.04	8	724.26	3.72
10	0.17	0.015	0.88	0.11	18.47	25	631.19	2.62
11	0.01	0.006	0.36	0.02	5.76	13	671.33	4.10
12	0.17	0.016	0.57	0.05	1.24	13	657.55	3.39
13	0.16	0.015	0.42	0.24	10.43	8	730.83	3.73
14	0.13	0.005	0.21	0.27	13.63	12	712.13	2.06
15	0.02	0.011	0.79	0.42	14.28	22	658.45	1.96
16	0.11	0.019	0.64	0.33	17.13	20	663.05	2.92
17	0.07	0.010	0.45	0.38	3.83	19	626.24	3.52
18	0.08	0.010	0.97	0.22	12.56	21	641.73	2.20
19	0.12	0.019	0.12	0.43	8.94	15	667.88	2.21
20	0.08	0.019	0.31	0.04	1.54	23	603.46	3.61
21	0.06	0.007	0.50	0.37	15.19	23	651.54	2.16
22	0.09	0.017	0.03	0.00	3.11	14	644.97	4.25
23	0.05	0.005	0.80	0.23	8.49	7	743.54	2.94
24	0.02	0.009	0.08	0.18	13.9	17	672.36	0.97
25	0.04	0.007	0.52	0.47	17.96	19	689.07	2.89
26	0.03	0.017	0.01	0.09	1.91	17	631.88	6.35
27	0.09	0.011	0.4	0.19	6.67	21	618.97	1.18
28	0.10	0.020	0.37	0.28	0.89	6	728.44	3.45
29	0.10	0.010	0.96	0.16	10.81	24	623.68	3.04
30	0.08	0.014	0.75	0.48	7.81	22	632.46	2.94
31	0.13	0.012	0.19	0.22	0.37	9	691.82	3.08
32	0.15	0.010	0.44	0.01	11.27	24	617.45	1.18
33	0.18	0.016	0.73	0.32	11.77	9	731.69	3.24
34	0.07	0.009	0.27	0.34	17.42	21	663.76	2.35
35	0.19	0.015	0.76	0.31	6.45	5	771.99	1.55
36	0.09	0.008	0.65	0.07	7.13	15	652.48	1.92
37	0.07	0.006	0.52	0.19	4.18	18	634.46	1.58
38	0.04	0.007	0.71	0.38	9.52	7	755.8	4.23
39	0.12	0.019	0.18	0.07	15.47	14	681.87	1.27
40	0.17	0.018	0.85	0.44	12.13	22	641.67	2.24

Notes: The best results are in bold.

Table W2: Experimental Parameter Combinations and Resulting Perplexity (Continued)

Comb.	α	γ	σ^2	a	b	K	$\bar{\mathcal{Y}}$	std(\mathcal{Y})
41	0.19	0.020	0.25	0.05	9.84	14	662.60	3.09
42	0.15	0.014	0.89	0.13	16.55	12	704.44	1.84
43	0.03	0.007	0.58	0.36	10.32	18	663.09	3.42
44	0.06	0.009	0.32	0.26	13.24	5	795.95	3.97
45	0.18	0.018	0.47	0.26	14.80	13	698.20	2.06
46	0.14	0.014	1.00	0.42	14.61	10	738.11	5.63
47	0.10	0.008	0.05	0.29	16.04	16	687.81	3.74
48	0.14	0.012	0.28	0.45	4.35	15	655.23	4.53
49	0.05	0.011	0.34	0.17	7.36	7	735.48	4.43
50	0.14	0.009	0.93	0.15	3.39	14	650.26	2.00
51	0.14	0.014	0.57	0.49	18.77	20	680.28	1.04
52	0.16	0.015	0.85	0.46	12.77	6	783.15	3.55
53	0.07	0.008	0.09	0.30	19.34	22	665.48	2.35
54	0.05	0.013	0.15	0.36	4.85	19	629.47	3.20
55	0.18	0.017	0.14	0.25	2.37	11	666.40	2.04
56	0.13	0.018	0.23	0.11	6.05	10	687.76	1.54
57	0.19	0.017	0.61	0.17	5.21	17	631.17	2.77
58	0.12	0.011	0.54	0.40	15.95	9	751.82	3.99
59	0.03	0.018	0.39	0.12	9.32	12	682.77	2.32
60	0.01	0.013	0.03	0.41	0.19	16	644.40	5.04

Notes: The best results are in bold.

E Substantive Outputs

Table W3: Topic Distributions of Sample Recall Statements

	Topic 10 Airbag	Topic 17 Power train	Topic 20 Electrical system	Topic 21 Vehicle speed control	Topic 22 Seat and seatbelt
Recall campaign: 00V228005	MMY: Nissan Quest 2000		Components: Seat belts		
Recall defect and consequence summary: Passenger vehicles equipped with TRW seat belt buckle assemblies. The buckle base of these seat belt assemblies were not properly heat treated and therefore do not pass the load bearing requirement of federal motor vehicle safety standard NO. 209, " seat belt assemblies ." In the event of a vehicle crash, the occupant may not be properly restrained.					
θ_i	0.267	0.001	0.001	0.001	0.236
$\sum_{j=1}^J \rho_{i,j} v_{i,j}$	0.018	0.000	0.000	0.000	0.250
Recall campaign: 18V404000	MMY: Mazda MPV 2005		Components: Air bags		
Recall defect and consequence summary: Mazda North America Operations (Mazda) is recalling certain 2005-2006 Mazda MPV vehicles sold, ... These vehicles are equipped with certain air bag inflators assembled as part of the passenger frontal air bag modules used as original equipment or replacement equipment. In the event of a crash necessitating deployment of the passenger frontal air bag , these inflators may explode due to propellant degradation occurring after long-term exposure to absolute humidity and temperature cycling. An inflator explosion may result in metal fragments striking the vehicle occupants resulting in serious injury or death.					
θ_i	0.147	0.068	0.071	0.018	0.082
$\sum_{j=1}^J \rho_{i,j} v_{i,j}$	0.162	0.038	0.087	0.049	0.002
Recall campaign: 09V388000	MMY: Toyota Camry 2009		Components: Vehicle speed control		
Recall defect and consequence summary: Toyota is recalling certain model year 2004-2010 passenger vehicles. The accelerator pedal can get stuck in the wide open position due to its being trapped by an unsecured or incompatible driver's floor mat. A stuck open accelerator pedal may result in very high vehicle speeds and make it difficult to stop the vehicle, which could cause a crash, serious injury or death.					
θ_i	0.005	0.177	0.049	0.172	0.000
$\sum_{j=1}^J \rho_{i,j} v_{i,j}$	0.066	0.079	0.070	0.145	0.000
Recall campaign: 14V355000	MMY: Buick LaCrosse 2008		Components: Electrical system		
Recall defect and consequence summary: Until this recall is performed, customers should remove all items from their key rings , leaving only the ignition key . The key fob (if applicable), should also be removed from the key ring . General Motors LLC (GM) is recalling certain model year 2005-2009 Buick LaCrosse, 2006-2011 Buick Lucerne, 2000-2005 Cadillac DeVille, 2006-2011 Cadillac DTS, 2006-2014 Chevrolet Impala, and 2006-2007 Chevrolet Monte Carlo vehicles. In the affected vehicles, the weight on the key ring and road conditions or some other jarring event may cause the ignition switch to move out of the run position, turning off the engine. If the key is not in the run position, the air bags may not deploy if the vehicle is involved in a crash, increasing the risk of injury. Additionally, a key knocked out of the run position will cause loss of engine power, power steering, and power braking, increasing the risk of a vehicle crash.					
θ_i	0.057	0.004	0.209	0.024	0.004
$\sum_{j=1}^J \rho_{i,j} v_{i,j}$	0.002	0.062	0.101	0.079	0.078
Recall campaign: 19V072000	MMMY: Mazda MX-5 Miata 2016		Components: Power train		
Recall defect and consequence summary: Mazda North American Operations (Mazda) is recalling certain 2016-2019 MX-5 (Miata) vehicles with automatic transmissions . Due to incorrect programming of the Transmission Control Module (TCM) , certain conditions may cause the vehicle to unexpectedly downshift and abruptly decelerate. If the vehicle abruptly downshifts, the driver may lose control of the vehicle, increasing the risk of a crash.					
θ_i	0.005	0.255	0.004	0.044	0.000
$\sum_{j=1}^J \rho_{i,j} v_{i,j}$	0.000	0.233	0.000	0.181	0.000

Table W4: Cosine Similarity Between Recall Statements and (Non-) Associated Complaints

Automaker	#Recalls	$Cos(\theta_i, v_i)$	$Cos(\theta_i, v_{n \neq i})$
Acura	34	0.333***	0.228
BMW	34	0.403***	0.231
Buick	22	0.414***	0.284
Cadillac	17	0.354***	0.240
Chevrolet	80	0.359***	0.260
Chrysler	24	0.375***	0.258
Dodge	31	0.347**	0.273
Ford	72	0.324***	0.257
GMC	15	0.422***	0.267
Honda	62	0.324***	0.257
Hyundai	71	0.351***	0.272
Jeep	27	0.324*	0.289
Kia	56	0.339***	0.269
Lexus	46	0.411***	0.241
Mazda	43	0.380***	0.240
Mercedes-Benz	43	0.390***	0.267
Nissan	68	0.383***	0.256
Subaru	31	0.356***	0.248
Toyota	84	0.370***	0.245
Volkswagen	36	0.386***	0.284

Note. For each automaker, we compute the cosine similarity of the topic distributions between a recall statement and its associated consumer complaints, denoted by $Cos(\theta_i, v_i)$, and its non-associated consumer complaints, denoted by $Cos(\theta_i, v_{n \neq i})$. We construct the non-associated complaints for each recalled MMY i as a set of complaints that are about the same automaker as MMY i but are not associated with any recall in our data. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

F Predicting Recall at the Current Time Period

We consider a set of metrics commonly used in the literature for evaluating the prediction performance across different model inputs under different types of prediction models. These metrics are defined based on the so-called *confusion matrix* (shown in Table W5): *sensitivity* ($= TP/(TP + FN)$), *specificity* ($= TN/(TN + FP)$), *precision* ($= TP/(TP + FP)$), *balanced accuracy* ($= (sensitivity + specificity) / 2$), and *F1 score* ($= 2 (precision \times sensitivity) / (precision + sensitivity)$). We also use the AUC, which reflects the extent to which a model can distinguish between the recalled and non-recalled classes. For all of these metrics, a higher score indicates better performance.

Table W5: Confusion Matrix for the Prediction of Recall Events

	Predicted Recall	Predicted Non-Recall
Actual Recall	True Positive (TP)	False Negative (FN)
Actual Non-Recall	False Positive (FP)	True Negative (TN)

Table W6: Comparison of Prediction Performance across Different Models and Inputs (%)

Model	Input	Sensitivity	Specificity	Balanced	F1 Score	AUC
XGBoost	HDPYP _p	73.73*	86.75*	80.24*	82.82*	88.08*
	HDPYP _{Equal}	70.59	80.72	75.66	79.82	83.24
	LDA _{Equal}	71.37	77.11	74.24	79.82	83.18
	Component	67.45	74.70	71.07	76.79	78.39
	n.a.	65.49	72.29	68.89	75.06	76.22
Logistic Reg.	HDPYP _p	71.76*	75.90*	73.83*	79.91*	81.21*
	HDPYP _{Equal}	68.63	73.49	71.06	77.43	77.31
	LDA _{Equal}	68.24	73.49	70.86	77.16	77.54
	Component	67.84	72.29	70.07	76.72	75.03
	n.a.	65.49	72.29	68.89	75.06	73.29
Random Forest	HDPYP _p	68.24*	74.70*	71.47*	77.33*	76.24*
	HDPYP _{Equal}	61.18	62.65	61.91	70.59	67.44
	LDA _{Equal}	55.69	66.27	60.98	66.82	65.28
	Component	60.78	63.86	62.32	70.45	69.23
	n.a.	58.43	67.47	62.95	69.14	68.16
Extended Cox	HDPYP _p	63.53*	61.45*	62.49*	72.16*	67.04*
	HDPYP _{Equal}	59.22	55.42	57.32	68.17	61.63
	LDA _{Equal}	58.04	55.42	56.73	67.27	60.84
	Component	57.65	54.22	55.93	66.82	61.83
	n.a.	59.22	57.83	58.52	68.48	60.37

Note. These input methods differ in how the textual information in consumer complaints is processed and aggregated to construct the covariates that enter a given prediction model. *The input method is the best at $p < 0.05$.

Table W6 compares the out-of-sample accuracy of predicting whether a MMY should be recalled now given all the information observed so far. We find that when the topic distributions of consumer

complaints are aggregated by $HDPYP_{\rho}$, the accuracy is significantly higher than the two benchmarks based on the arithmetic mean (i.e., $HDPYP_{Equal}$ and LDA_{Equal}) across all of the metrics and all of the model specifications ($p < 0.05$). In particular, under the XGBoost which performs the best among the four prediction models, the $HDPYP_{\rho}$ -based model inputs improve the accuracy by 3% to 13%, compared with the LDA_{Equal} -based model inputs. It seems that the topic distributions estimated by the HDPYP itself have slightly better predictability than the topic distributions estimated by the LDA, but the most improvement is from properly weighting the information using ρ . In addition, we find that simply averaging the topic distributions of the consumer complaints estimated by the LDA does not necessarily improve the prediction, compared to not including the topic distributions at all. Taken together, our findings suggest that the value of the textual information in consumer complaints can be realized only when the few important complaints are identified, further confirming the importance of mining consumer complaints using the HDPYP. Finally, compared with the other prediction models, the XGBoost model seems to achieve a better balance between detecting situations that trigger a recall (i.e., sensitivity) and detecting those that do not (i.e., specificity).

G Drafting Recall Statements

Table W7: Auto-Generated Recall Statements Using Different Inputs - III

MMY: Acura RSX 2003	Recall campaign: 06E023000	Components: Exterior lighting
Recall defect and consequence summary:		
Certain Anzo combination headlamp assemblies sold as replacement lamps for use on certain passenger vehicles. Some combination lamps are not equipped with amber side reflectors which fail to conform to federal motor vehicle safety standard No. 108, "lamps, reflective devices, and associated equipment." without the amber reflectors, the vehicle will be poorly illuminated, possibly resulting in a vehicle crash.		
Top-10 complaints ranked by HDPYP:		
Certain Anzo combination headlamp assemblies sold as replacement lamps for use on certain passenger vehicles. Some combination lamps are not equipped with amber side reflectors which fail to conform to federal motor vehicle safety standard No. 108, lamps, reflective devices, and associated equipment. Without the amber reflectors, the vehicle will be poorly illuminated, possibly resulting in a vehicle crash without warning.		
Random 10 complaints:		
On certain passenger vehicles equipped with automatic transmissions, the fuel pump hose may come out of its mounting position. If the hose comes out of its mounting position, the fuel pump could leak fluid. This could cause a fire to occur without warning.		
Complaints with fired, crashed, injured or death:		
On certain passenger vehicles, the front suspension lower struts may break due to improper welding. This could result in the strut breaking and increasing the risk of a crash.		
All complaints:		
Honda (American Honda Motor Co.) is recalling certain model year 2003-2004 Acura RSX and RSX hybrid vehicles manufactured October 1, 2004, to January 31, 2004, and 2005-2006 Honda GR-V vehicles manufactured February 21, 2004, to July 31, 2006. The affected vehicles have a fuel leak inlet that may result in a fuel leak. A fuel leak in the presence of an ignition source can result in a fire.		

Table W8: Auto-Generated Recall Statements Using Different Inputs - IV

MMY: Toyota Matrix 2005	Recall campaign: 06E023000	Components: Exterior lighting
Recall defect and consequence summary:		
Certain Anzo combination headlamp assemblies sold as replacement lamps for use on certain passenger vehicles. Some combination lamps are not equipped with amber side reflectors which fail to conform to federal motor vehicle safety standard No. 108, "lamps, reflective devices, and associated equipment." without the amber reflectors, the vehicle will be poorly illuminated, possibly resulting in a vehicle crash.		
Top-10 complaints ranked by HDPYP:		
Certain pro-a motors corner lamps, turn signals, and headlights sold as replacement lamps for use on certain passenger vehicles listed above. Some combination lamps that are not equipped with amber side reflectors fail to conform to federal motor vehicle safety standard No. 108, lamps, reflective devices, and associated equipment. Without the amber reflectors, the vehicle will be poorly illuminated, possibly resulting in a vehicle crash without warning.		
Random 10 complaints:		
Toyota is recalling certain model year 2005-2010 passenger vehicles for failing to comply with the requirements of federal motor vehicle safety standard No. 110, "tire selection and rims." these vehicles were sold without the requisite load carrying capacity modification labels. A driver may overload a vehicle which could increase the risk of a crash.		
All complaints:		
On certain sport utility vehicles, due to a manufacturing issue in the front suspension lower ball joint, there is an increased risk of a crash.		

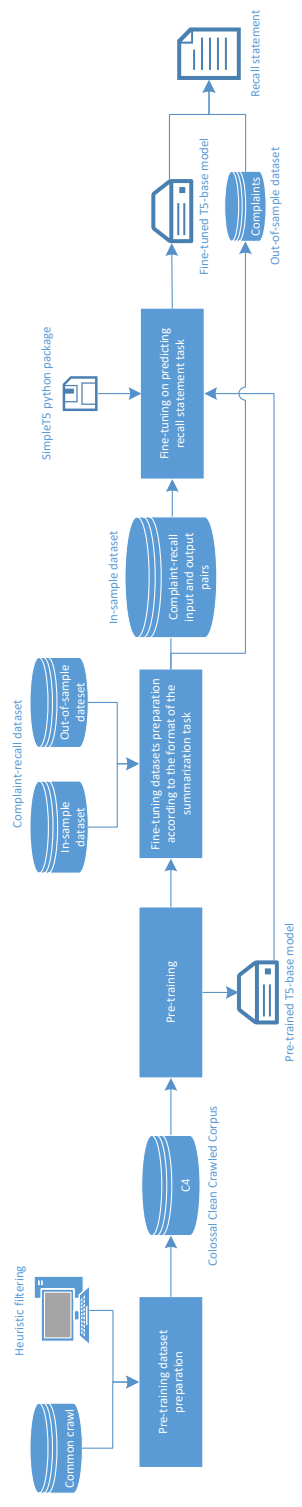


Figure W3: The Flow of Using Pre-trained and Fine-tuned T5 to Generate Recall Statements

References

- Ankenman, B. E., B. L. Nelson, and J. Staum (2010). Stochastic Kriging for simulation metamodeling. *Operations Research* 58(2), 371–382.
- Buntine, W. and M. Hutter (2012). A Bayesian view of the Poisson-Dirichlet process. *ArXiv preprint arXiv: 1007.0296v2*.
- Chen, C., L. Du, and W. Buntine (2011). Sampling table configurations for the hierarchical Poisson-Dirichlet process. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis (Eds.), *Machine Learning and Knowledge Discovery in Databases*, Berlin, Heidelberg, pp. 296–311. Springer Berlin Heidelberg.
- Du, L., W. Buntine, and H. Jin (2012). Modelling sequential text with an adaptive topic model. In Tsujii, J., Henderson, J., and Pasca, M. (Ed.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 535–545. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics.
- Kleijnen, J. P. C. (2015). *Design and Analysis of Simulation Experiments* (2 ed.). New York: Springer.
- Lim, K. W., W. Buntine, C. Chen, and L. Du (2016). Nonparametric Bayesian topic modelling with the hierarchical Pitman–Yor processes. *International Journal of Approximate Reasoning* 78, 172–191.
- Loeppky, J. L., J. Sacks, and W. J. Welch (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics* 51(4), 366–376.
- Mochihashi, D. and E. Sumita (2008). The infinite Markov model. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Eds.), *Advances in Neural Information Processing Systems 20*, pp. 1017–1024. MIT Press, Cambridge, MA: Curran Associates, Inc.
- Oldham, K. B., J. Myland, and J. Spanier (2009). *An Atlas of Functions: With Equator, the Atlas Function Calculator the Atlas Function Calculator*. New York, USA: Springer Science and Business Media, New York City.
- Puranam, D., V. Narayan, and V. Kadiyali (2017). The effect of calorie posting regulation on consumer opinion: A flexible latent dirichlet allocation model with informative priors. *Marketing Science* 36(5), 726–746.

- Salemi, P., J. Staum, and B. L. Nelson (2019). Generalized integrated Brownian fields for simulation metamodeling. *Operations Research* 67(3), 874–891.
- Santner, T. J., B. J. Williams, and W. I. Notz (2018). *The Design and Analysis of Computer Experiments* (2 ed.). New York: Springer.