Communications of the Association for Information Systems

Volume 47 Article 16

11-9-2020

Evaluating Topic Modeling Interpretability Using Topic Labeled Gold-standard Sets

Biagio Palese
Northern Illinois University, bpalese@niu.edu

Gabriele Piccoli
Louisiana State University, misqe_eic@aisnet.org

Recommended Citation

Palese, B., & Piccoli, G. (2020). Evaluating Topic Modeling Interpretability Using Topic Labeled Gold-standard Sets. Communications of the Association for Information Systems, 47, pp-pp. https://doi.org/10.17705/1CAIS.04720

Communications of the Association for Information Systems

Research Paper DOI: 10.17705/1CAIS.04720 ISSN: 1529-3181

Evaluating Topic Modeling Interpretability Using Topic Labeled Gold-standard Sets

Biagio Palese
Operations Management and Information Systems
Northern Illinois University
bpalese@niu.edu

Gabriele Piccoli
Stephenson Department of Entrepreneurship & Information Systems
Louisiana State University

Abstract:

The paucity of rigorous evaluation measures undermines topic modeling results' validity and trustworthiness. Accordingly, we propose a method that researchers can use to select models when they assess topics' human interpretability. We show how they can evaluate different topic models using gold-standard sets that humans label. Our approach ensures that the topics extracted algorithmically from an entire corpus concur with the themes humans would have identified in the same documents. By doing so, we combine human coding's advantages for topic interpretability with algorithmic topic Modeling's analytical efficiency and scalability. We demonstrate that one can rigorously identify optimal model parametrizations for maximum interpretability and to rigorously justify model selection. We also contribute three open access gold-standard sets in the hospitality context and make them available so other researchers can use them to benchmark their models or validate their results. Finally, we showcase a methodology for designing and developing gold-standard sets for validating topic models, which researchers interested in developing gold-standard sets in domains and contexts appropriate for their research can use.

Keywords: Human Interpretable Topics, Gold-standard Set, Text Mining, Topic Evaluation, Topic Interpretability Measure, Topic Modeling.

This manuscript underwent editorial review. It was received 02/26/2019 and was with the authors for thirteen months for three revisions. Iris Junglas served as Associate Editor.

1 Introduction

Unstructured text data continues to increase in amount at a tremendous pace. People communicate by sharing posts on social media, organizations use chat bots for live customer service calls, smart speakers come with increasing functionality for voice interaction (so-called skills), and online review platforms provide millions of user-generated opinions. For example, the number of reviews on TripAdvisor more than doubled between 2014 and 2016 from 200 million to 465 million (TripAdvisor, 2017), and, in the fourth quarter of 2018, the website contained over 730 million reviews (TripAdvisor, 2019).

As the amount of private and publicly available textual data increases, both academic researchers and commercial ventures have the opportunity to extract valuable information. Text mining offers a solution. It transforms text data into information when human coders do not constitute a viable option (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010; Debortoli, Müller, Junglas, & vom Brocke, 2016). As a consequence, algorithmic methods to analyze text data, such as sentiment analysis and topic modeling, have become increasingly important methodologies in information systems (IS) and related disciplines. Recently, Eickhoff and Neuss (2017) analyzed journals in the Financial Times 50 ranking from 1975 to 2017 and found that 145 IS papers adopted topic modeling. However, they found that only 36 percent of such papers validated the models they used compared to 59 percent of the papers in other disciplines (Eickhoff & Neuss, 2017).

While the methodology has gained acceptance, criticism lingers. The main concern relates to selecting and evaluating models (Nikolenko, Koltcov, & Koltsova, 2017). For example, only six percent of the 145 IS papers that Eickhoff and Neuss (2017) identified developed or compared assessment approaches to validate topic models. Moreover, different disciplines have approached the assessment problem from different perspectives. In machine learning and computer science, researchers focus on assessing models' predictive ability and on identifying the optimal number of topics to extract (Wallach, Murray, Salakhutdinov, & Mimno, 2009). In computational linguistics, researchers focus on topics' semantic meaning (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009) and their internal coherence (Newman, Lau, Grieser, & Baldwin, 2010; Mimno, Wallach, Talley, Leenders, & McCallum, 2011). In other areas, such as IS, researchers use topic models as an analytical methodology to extract and measure constructs in large corpora. For example, Palese and Usai (2018) extracted the five dimensions of service quality from online reviews to measure customer satisfaction. In these disciplines, topic modeling holds great potential for providing new construct-measurement and information-extraction opportunities (Müller, Junglas, Brocke, & Debortoli, 2016). For this reason, researchers need to be able to assess the extracted topics' human interpretability. Practicing managers in organizations that want to gain insight into how customers perceive customer service interaction logs or reviews also need to be able to do so.

Thus, a critical question remains unanswered: how can we measure the human interpretability of the thematic structures that topic modeling returns (Blei, 2012)? More broadly, how can researchers and reviewers know that their models extract meaningful topics from large corpora that, due to their dimensions, humans find inscrutable (Eickhoff & Neuss, 2017)?

In this paper, we offer a method for evaluating topics' interpretability based on human judgment. We create three human-labeled gold-standard sets of service quality dimensions for hotel reviews and make them available to the community. We use the term gold-standard set to describe a small collection of documents that we extracted from a dataset and that humans manually labeled. We use such a gold-standard dataset to establish a "ground truth" against which researchers can benchmark the results that algorithmic topic models generate. They can then use the gold-standard sets to benchmark different/new models. We also show how researchers can use the gold-standard sets to select models. Our work provides three main contributions. First, it demonstrates a method for selecting and evaluating models that focuses on the extracted topics' human interpretability. Second, it contributes three gold-standard sets (called inclusive, full agreement, and partial agreement) other researchers can use to benchmark their models or validate their results in the context of service quality in the hospitality industry. Third, it showcases a methodology for designing and validating gold-standard sets of text corpora. This approach enables researchers to create gold-standard sets specific to their context. Researchers can use the methodology in IS, marketing, hospitality, or any other discipline in which they need to reliably extract meaning from textual service data (e.g., transcripts of service chatbot conversations, social media service interactions, help desk support calls).

The paper proceeds as follows: in Section 2, we summarize previous work focused on selecting and evaluating topic models. In Section 3, we describe the sequential steps for creating a gold-standard set. In

Section 4, we illustrate the proposed method and demonstrate how to assess topics' human interpretability. In Section 5, we offer guidelines and recommendations for using our approach to validate topic models results. Finally, in Section 6, we conclude the paper.

2 Research on Selecting and Evaluating Topic Models

Topic modeling is a text-mining technique that enables one to extract large text corpora's thematic structure at a scale that humans find inscrutable. Topic models are probabilistic "latent variable models of documents that exploit the correlations among the words and latent semantic themes" (Blei & Lafferty, 2007, p. 18). They are generative probabilistic techniques that operate on the bag-of-words assumption (Blei, 2012). They model thematic text structures (i.e., topics) through distributions over words and documents as distribution over topics. Thus, a topic model generates a probabilistic model that captures (and can reproduce) topic structure that a set of documents implicitly contains.

Using topic modeling to analyze text allows one to inspect and "understand" topics in terms of the distribution of terms that comprise them. For each document, the topic model produces a vector of weights that represent how strongly each topic pertains to the document. However, given the many degrees of freedom in choosing parametrizations, researchers often find selecting the appropriate topic model to represent a corpus challenging. While evaluating a topic model's predictive ability has attracted most research attention to date, topics' interpretability remains a central challenge for scholars and managers (Chang et al., 2009).

Recently, multiple studies have focused on finding the best way to select and evaluate competing models' quality. We argue that what metric one selects depends on what type of topic model one uses (unsupervised vs. weakly supervised) and one's analysis objective. An exploratory analysis that one designs to investigate the content of multiple datasets requires different validation priorities than a project that one designs to inform a firm about an important business decision. For this reason, researchers need to select the metric that aligns best with their research's aim and use it to compare multiple models to identify the best performing one (Doshi-Velez & Kim, 2017). Only by assessing models through metrics can researchers validate their selection of a specific model's parametrization for data analysis. Moreover, by showing that they have evaluated their models' results, researchers will increase readers' faith in their work's rigor.

2.1 LDA: Seeding and Parametrization

Many different topic modeling algorithms exist, but, in this paper, we focus on the most widely adopted one; that is, latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003). LDA is easily described by its imaginary generative process, which assumes that documents in a corpus are collections of a fixed set of topics. Each topic is characterized by a distribution of words confined to the vocabulary of the corpus. Topics are then assigned to each document in a different proportion. The LDA algorithm estimates the hidden topics in each document and the word distributions in each topic given how often the observed words occur in each document. When beginning a research project, researchers need to decide whether to use the algorithm to explore the number of topics or to specify a priori the number and meaning of the topics that comprise the documents in their corpus. Researchers should use the first approach, called unsupervised topic modeling, when they want to summarize a large set of documents and extract their hidden thematic structure without guiding the search for specific topics. They should use the second one, called weakly supervised (or seeded) topic modeling, when they want to examine the distribution of pre-selected topics in a collection of documents, such as when they have a priori knowledge of the documents' content or when strong theory guides their efforts to search and extract topics that represent specific constructs. In seeded topic modeling, researchers focus on categorizing documents based on the probability they address known topics or topics that they expect the documents to discuss and on treating any other text as irrelevant by creating an "undefined" topic. The term weakly supervised refers to the fact that the topic modeling algorithm will start with a list of terms (i.e., the seeds) the researcher knows to characterize each topic. For example, when seeding a topic such as "hotel location", one may start with terms like "place," "view", and "transportation". Thus, while unsupervised topic modeling does not require any a priori knowledge, the weakly supervised approach demands that researchers specify the number of topics and select topic-specific seed words (Lu, Ott, Cardie & Tsou, 2011). Using this a priori knowledge, the LDA algorithm will create the term distribution in the topics and the topic distribution in the documents. In this paper, we focus on weakly supervised topic modeling because this approach mirrors the job that human raters perform in manually labeling documents based on topics of interest.

Given that the number of topics researchers need to extract depends on a priori research decisions, the remaining major choices for configuring a topic model pertain to the model's parametrization (specifically, the alpha and delta hyperparameters). The former affects the topic proportion in the documents, while the latter impacts the distribution of terms in each topic. Specifically, smaller alpha values affect the number of topics by forcing the topic modeling algorithm to spread the topic distributions over only a few topics for each document. Smaller delta values affect the distribution of words in each topic by concentrating the word probability on a few characterizing terms for each topic.

2.2 Existing Evaluation Metrics

The literature has advanced various evaluation metrics to assess topic models. Many focus on models' performance when predicting unseen data. For this reason, they optimally suit exploratory research in which researchers focus on determining the appropriate number of topics to extract (Blei & Lafferty, 2007). Among them, held-out likelihood measures (e.g., perplexity) have gained consensus in the literature. These measures require training the topic model on a corpus subset so different parametrizations can be evaluated on a test set. In this case, researchers focus on measuring how well the model represents or reproduces the statistics of the held-out (test) data. However, these purely statistical approaches provide no information about the content of the extracted topics (Boyd-Graber, Mimno, & Newman, 2014). Moreover, they say nothing about the topics' accuracy and interpretability in relation to what human raters would identify. Furthermore, topics' interpretability has an inverse relationship with the model's predictive ability (Chang et al., 2009).

Other measures evaluate the quality of the topics that one extracts from a corpus (Chang et al., 2009) to validate the terms included in each topic (so-called word intrusion) and assess each topic's semantic meaning (so-called topic intrusion). Word intrusion "measures how semantically cohesive the topics inferred by a model are and tests whether topics correspond to natural grouping for humans" (Chang et al., 2009, p. 2). Topic intrusion "measures how well a topic model's decomposition of a document as a mixture of topics agrees with human association of topics with a document" (Chang et al., 2009, p. 2). While these measures allow one to evaluate topics' interpretability, one can perform them only on a small subset of documents. Moreover, they require much time and money to perform because they rely on humans to assess the corpus under investigation (Boyd-Graber et al., 2014).

Other measures focus on topics' semantic coherence. They focus on identifying which thematic structures the topic model extracts are random or illogical. Once the measures identify illogical structures, they remove them from the analysis. These measures rest on the assumption that topics display coherence if pairs of their top terms have a high degree of association (Newman, Noh, Talley, Karimi, & Baldwin, 2010). These measures produce a topic coherence score "based on a pointwise mutual information of pairs of terms taken from topics" (Boyd-Graber et al., 2014 p. 241). The topic coherence score assesses the word association between pairs of frequent words in each topic. High scores suggest the topic displays semantic coherence. Research shows that semantic coherence measures correlate with human-judged topic coherence (Mimno et al., 2011; Newman et al., 2010). However, they consider only the top terms in each topic and do not evaluate the topic's assignment to the document in the corpus. In Table 1, we summarize the metrics we describe above versus our approach, which we call human interpretable topics (HIT).

Table 1. Evaluation Metrics Comparison

	Evaluation level	Evaluation method	Topic interpretability
Held-out likelihood	Corpus	Algorithmic	No
Word intrusion	Individual topic	Human assessed on a corpus subset	Yes
Topic intrusion	Set of topics	Human assessed on a corpus subset	Yes
Coherence score	Individual topic	Algorithmic	No
HIT (new)	Corpus	Algorithmic	Yes

3 Our Proposed Solution: Human Interpretable Topics (HIT)

We designed our proposed solution, which we call HIT, to combine the advantages that both human assessment and algorithmic scalability provide. HIT measures the degree to which humans can interpret extracted topics. In fact, people evaluate the topics against a gold-standard set of manually labeled documents. While the gold-standard set requires humans to assess and identify the topics in the sample documents (see below), the approach is algorithmic and scalable. Moreover, researcher and practitioners who use a published gold-standard set such as the one we provide will need no manual human intervention. Thus, once researchers produce the gold-standard set, they can generalize and scale model validation. This approach directly responds to calls to measure how closely topic modeling results approach human topic assignments (Chang et al., 2009). Such an assessment will enable researchers to show, using real documents from the corpus that they investigate, that the topic model they fit meaningfully represents the themes in the corpus. Here, "meaningful" refers a consensus in how human raters assign topics. In other words, our approach focuses on ensuring the topics that one extracts algorithmically from a corpus match the topics human judges would have manually extracted from the same corpus.

In Figure 1, we outline the pipeline for extracting human-interpretable topics. In the method we propose, researchers first create a gold-standard set that includes manually labeled predefined topics of interest for their study. Researchers use the gold-standard set to benchmark the results from different classification and clustering models. For example, a recent study in IS used a human-labeled gold-standard set with annotated sentiment polarity to benchmark 20 sentiment-analysis tools (Abbasi, Hassan, & Dhar, 2014). Once researchers have a gold-standard set specific to their study's context, they can run topic modeling on the corpus. The corpus should not contain documents that researchers include in the gold-standard set. To identify the topic modeling's optimal parametrization, researchers run multiple models with different hyperparameters. They then use the fitted models to classify the documents in the gold-standard set. At this point, they compare each model's results against the gold-standard set that humans manually labeled. They compute each model's accuracy by measuring discrepancies between the algorithmic-labeled and the human-labeled gold-standard set. HIT improves on existing methods in that, once researchers have produced the gold-standard set, they can generalize and scale model validation. Thus, they benchmark each model's parametrization results against the gold-standard sets that human raters create to compare the results from different model parametrizations. The model with the highest accuracy represents the optimal model. Humans can interpret the optimal model results because they match humans' judgments. Researchers can then select and use this model for their analyses.

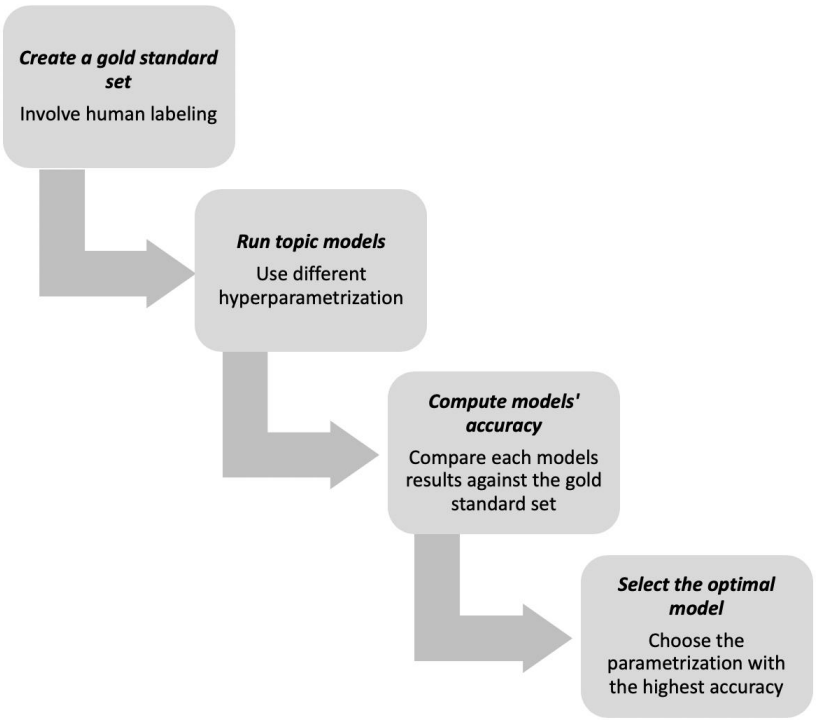

Figure 1. Human Interpretable Topics Roadmap

In Section 3.1 to 4.2, we illustrate the HIT pipeline with an example from the hospital industry. In particular, we show how we used HIT to classify a corpus focused on service quality in hotels. Online reviews and corresponding managerial responses constituted the unit of analysis for our classification. To the best of our knowledge, no publicly available labeled dataset for service quality in the hospitality industry exists. Thus, we first created the gold-standard set. In this way, we contribute the first open-access gold-standard sets in the hospitality context that researchers can use.

3.1 Creating a Human-labeled Gold-standard Set for Service Quality in the Hotel Industry

To create a human-labeled gold-standard set for service quality in the hotel industry, one needs to tag each online review or response an arbitrary number of topics—from zero to (theoretically) infinity. Thus, tagging these documents[1] involves much more complexity than a typical machine-learning tagging exercise, such as creating an image dataset such as Caltech101 (Fei-Fei Li & Perona, 2005). In a typical image-tagging exercise, raters simply have to identify the object or objects in a picture. Researchers can crowdsource this task as many have done with the reCAPTCHA process (e.g., find all images with a storefront in it). Conversely, a gold-standard set for topic modeling in a specific domain (such as hotel service) requires researchers to extensively train the raters so they clearly understand each relevant topic's definition. Moreover, raters can best identify topics at the document level because interpreting each topic relies on contextual information that may exist in the previous or following sentences. Therefore, raters can possibly see different predefined topics for the same document. For this reason, we produced three gold-standard sets: inclusive, full agreement, and partial agreement. The inclusive set contained all the documents and any topics that any rater identified. Thus, it comprehensively included all topics that all labelers identified in all reviews and all responses. The full agreement set included the documents for which all raters identified the same topics. Finally, the partial agreement set contained documents for which all raters identified at least one common topic. For each document, review, or response, the partial agreement set included only those topics that all raters labeled. For example, in a situation with five raters, if each rater identified five distinct topics in a document but only two topics appeared across all five raters, then the partial agreement set for this document would show only those two topics. In Figure 2, we lay out the steps necessary to create context-specific gold-standard sets.

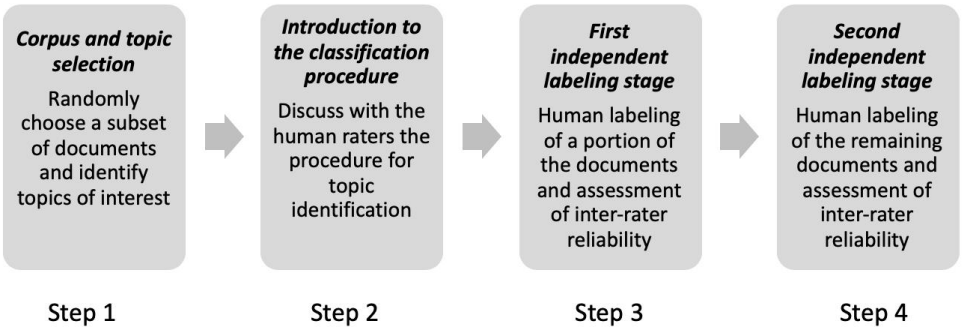

Figure 2. Steps to Create a Gold-standard Set

3.1.1 Step 1: Corpus and Topic Selection

In the first step, we selected a dataset of online reviews and responses from which to extract the gold-standard set. To do so, we obtained 43,981 reviews and responses from TripAdvisor, the dominant online review system for collecting and sharing hotel guests' opinions. The review addressed 1,764 different hotels in the Unites States and captured all hotels in the 25 most populous cities in the country. From the set, we extracted 500 randomly selected documents (250 reviews and 250 responses).

We manually labeled each document in the gold-standard set using one of seven predefined topics. The first five topics concerned the five elements of the lodging experience (i.e., service, value, location, room, and food), which guests commonly use to evaluate their experience in hotels (Piccoli & Ott, 2014; Piccoli, 2016). To these five topics, we added hotel amenities, which empirically emerged from preliminary

[1] For the rest of the paper we use the term documents when we want to refer to the reviews and responses available in our corpus

analyses, and greetings and salutations, which generally appear in responses (we define and provide examples of each predefined topic in Appendix A). These predefined topics offer information not only about the service encounter (via the predefined topics service, room, food, and greeting and salutations) but also about a hotel's physical characteristics (via the predefined topics location and hotel amenities) and about the overall customer experience (via the predefined topic value).

3.1.2 Step 2: Introduction to the Classification Procedure

In the second step, we selected four online review system users (two males and two females) to manually label the gold-standard set documents. The lead researcher organized a meeting with all the validators[2]. During the meeting, the group reviewed the classification procedure (see Appendix A). The lead researcher taught the validators what the seven predefined topics meant and discussed real reviews and responses as exemplars with them. At the meeting, the lead researcher answered questions and offered clarification about the definitions and examples.

After the training, the five raters and the lead researcher independently classified an initial batch that comprised 20 documents. They could assign as many topics as they identified in each document. We instructed the raters to tag a document as unclassifiable when they did not recognize any of the seven topics. After the raters completed the categorization, the lead researcher then discussed each document individually to make sure the labelers understood the topics clearly. The researcher further discussed and explained the topics to resolve disagreements. However, we did not modify the original labeling when the discussion pointed toward disagreement. Before moving on to the third step, the lead researcher computed inter-rater reliability by calculating Fleiss' kappa, a standard of agreement among multiple raters for categorical labels (Landis & Koch, 1977) for each step in the classification process. Kappa values range from 0 to 1 with values above 0.61 indicating substantial inter-rater reliability and those above 0.81 indicating almost perfect inter-rater reliability. The Fleiss' kappa value for the inclusive set was 0.90, which indicates almost unanimous consensus. By definition, the full and partial agreement sets had Fleiss' kappa of 1 because they included only documents in which all raters identified the same topics. For this reason, at each step in the classification procedure, we report the Fleiss' kappa value only for the inclusive set.

3.1.3 Step 3: First Independent Labeling Stage

In this step, all five raters individually labeled another 80 documents without the lead researcher's supervision. The researcher instructed the raters to use their judgment to classify content that straddled different topics given the definitions that the classification procedure provided and to bring these issues to the next clarification meeting.

In the meeting, the lead researcher and raters resolved inconsistencies, and the researcher further trained the raters. The labelers discussed the topics they had identified for each document and raised issues emerging during the classification. The dialogue led to some changes in the original predefined topics' definitions to create clearer boundaries between the predefined topics. After the discussion and given the definition changes, the labelers revised their initial classification. The Fleiss' kappa value for the inclusive set at the end of this step was 0.90.

3.1.4 Step 4: Second Independent Labeling Stage

In the last step, the lead researcher asked the raters to classify the remaining 400 documents without supervision and to bring their results to a final consolidation meeting. During the meeting, the five raters discussed all the documents close to full consensus, including situations in which four out of the five labelers identified a topic and situations in which only one labeler categorized a topic. The raters could change the original classification if, and only if, they completely agreed with the change during discussion. At the end of this step, the Fleiss' kappa for the inclusive set was 0.87, which indicates almost perfect agreement. The full agreement set comprised 280 documents and the partial agreement set comprised 499 documents (only one document had zero common topics across validators).

[2] We use the terms validator, labeler, and rater interchangeably in this section.

Table 2. Gold-standard Sets Descriptive

	Inclusive set	Full agreement set	Partial agreement set
Total number of documents	500	280	499
Total number of topics	1,695	790	1,363
Average number of topics per document	3.39	2.82	2.73
Fleiss' kappa	0.87	1	1

3.2 Classification Results

At the end of the classification procedure, we isolated three different gold-standard sets that represented the raters' labeling in three distinct ways (see Table 2). The number of documents and topics in each set resulted from their different designs and purposes. On average, the partial agreement set had the lowest number of topics because it contained only topics that all the raters identified for each document. So, while the three sets offer flexibility to researchers in terms of the type of assessment they want to perform, researchers should not ignore their intrinsic differences when interpreting HIT results. For this reason, when evaluating the results, researchers also need to consider the number and proportion of predefined topics in each set. Table 3 reports the total number of predefined topics (count) and the proportion of each predefined topic relative to the total number of predefined topics (topics %) and the total number of documents (documents %) for the three sets.

Table 3. Predefined Topics Descriptive

Predefined topics	Inclusive set			Full agreement set			Partial agreement set		
	Count	Topics %	Documents %	Count	Topics %	Documents %	Count	Topics %	Documents %
Value	280	16.52	56.00	128	16.20	49.23	213	15.63	42.68
Service	340	20.06	68.00	170	21.52	60.71	277	20.32	55.51
Room	261	15.40	52.20	117	14.81	41.78	228	16.73	45.69
Location	251	14.81	50.20	115	14.56	41.07	215	15.77	43.09
Hotel amenities	193	11.39	38.60	66	8.35	23.57	113	8.29	22.64
Greetings and salutations	255	15.04	51.00	158	20.00	56.43	246	18.05	49.30
Food	115	6.78	23.00	36	4.56	12.86	71	5.21	14.23

Table 3 shows the predefined topics lacked homogeneity in each set. Some results were consistent among the sets; for instance, raters identified service the most often and food the least often. However, some predefined topics showed significant differences in the topics and document percentages. For example, the percentages for value, service, hotel amenities, and food dropped significantly from the inclusive set to the full agreement and partial agreement sets. Moreover, food and hotel amenities also dropped significantly in terms of topic percentage when comparing the inclusive set with the full and partial agreement sets. Thus, it seems some predefined topics have more volatility than others (food and hotel amenities in particular). To confirm this intuition, we looked at agreement among the raters at the topic level. Table 4 shows that service, room, location, and greeting and salutations reached agreement scores above 90 percent, which indicates that the raters consistently classified these predefined topics in the documents. On the other hand, the other predefined topics had lower agreement scores (hotel amenities in particular), which indicates the raters did not as consistently classify them.

A possible explanation for why the raters did not as consistently define these topics and hotel amenities in particular is that they straddle two topics, which makes it difficult for people to correctly classify them. For example, when guests describe their experience with the valet parking service (e.g., valet parking was slow), service would represent the correct label. However, if they simply mention the availability of valet parking, hotel amenities would represent the correct label (e.g., this hotel offers valet parking). This example shows the inherent difficulty that human raters experience in correctly classifying the predefined topics. However, we do not need to further investigate the documents that the raters classified in any

topics since the raters displayed high agreement overall. We now illustrate how topic modeling performs compared to human labeling.

Table 4. Per Topic Agreement Between Raters

Predefined topics	Agreement
Value	88.79%
Service	91.00%
Room	93.56%
Location	92.59%
Hotel amenities	78.34%
Greetings and salutations	98.59%
Food	81.22%

4 An Illustration: Extracting Human Interpretable Topics from Text

In this section, we demonstrate how researchers can use HIT to assess the interpretability of the topics in their corpora and how they can justify the topic model parametrization they select using one or more of the gold-standard sets that we provide.

Topic modeling constructs every document as a mixture of topics. We used weakly supervised topic modeling through Gibbs-sampling to extract the topics that the human labelers identified. For this illustration, we used multiple topic models with different parametrizations by varying the values of the hyperparameters alpha and delta. For illustrative purposes, we selected a wide range of values for each hyperparameter with small increments (a total of 150 different models). We considered 150 models as a sufficient number of models for this demonstration, but the total number of distinct models that one selects should depend on how much computational power one has and how fine-grained one needs to make one's analysis. For each of the 150 models, we extracted eight topics, which included the seven predefined topics that the human raters identified and one unsupervised topic to collect possible noise (or a potentially relevant topic we missed) in the corpus. Finally, to avoid overfitting, we excluded the gold-standard set documents from the corpus (since we used them as a holdout benchmark set).

4.1 Dataset and Topic Models' Parametrizations

After we set aside the elements of the gold-standard sets, the dataset contained 87,462 documents that comprised online reviews and their responses. We used a seeded sentence-level LDA model (Lu et al., 2011) that assigns one topic to each sentence. This approach treats each sentence as a document and takes the topic with the highest probability in that document as representing the sentence's label. In the case where two or more topics have the same probability, we labeled the sentence as undefined since no topic prevailed. We used this approach because it mirrors the human labeling process. By doing so, we identified a topic for each sentence before aggregating sentences to the review level.

Thus, in the first step in our analysis, we split the documents into sentences. We used the tokenizers package in R (Mullen, Keyes, Selivanov, Arnold, & Benoit, 2018) to complete this task, which yielded 669,037 sentences in total. We followed best practices on preprocessing the corpus using the tm package (Feinerer, Hornik, & Meyer, 2008; Feinerer & Hornik, 2018). We removed stop words, numbers, punctuation, and words with fewer than three and more than 30 characters. Then, we created the document term matrix (DTM). Before running the topic models, we removed from the DTM terms that appeared in less than 50 documents and documents that did not contain any terms after completing the preprocessing. The resulting document term matrix dimensions spanned 664,853 documents by 12,524 terms. Since we used weakly supervised topic modeling, we next specified the set of terms, known as seed words, that represent each topic's essence. We analyzed the corpus and selected six to seven words to seed each topic (Appendix B). We recommend selecting frequent words that are orthogonal among topics so they effectively differentiate the topics during the extraction. We then created a seed words matrix that included the topics we wanted to extract (rows) and all the terms available in our DTM (columns). Next, we focused on selecting the a priori weight of the selected seed words before fitting the topic models. Previous studies suggest using a small percentage of the total number of documents as the seed weight (Lu et al., 2011). Thus, we performed tuning between one and 10 percent to determine the

optimal percentage. We found three percent to be the optimal seed weight, so, for each seed word, we added 2,623.56 (3% of the total number of documents available in our dataset) as the a priori seed weight. We used the LDA function in the topicmodels R package (Grün & Hornik, 2011, 2018) to run the 150 topic models with the same seed words and seed weights. All the models also had the same specification in the LDA function for the number of iterations (1000), the number of burn-in iterations (500), and the number of thinning iterations (100). However, each model took a different combination of the hyperparameters alpha and delta. The alpha values ranged between 0.001 to 1.5, while the delta values ranged between 0.1 and 15. The values resulted from various tuning iterations that we conducted to demonstrate how they affected different models' performance and, thus, what one we selected as the best performing one. For each fitted model, we used their posterior probabilities to label the topics for the documents available in the gold-standard sets (benchmarking held-out sets). We used the posterior function in the topicmodels package to determine the posterior probabilities of the documents in the gold-standard set. By doing so, we had 150 different versions (one for each model) of the gold-standard sets that topic modeling labeled.

We then aggregated the sentence level topic assignments to the review or response level and computed the discrepancies between the topics that each model identified and the topics that the humans in the gold-standards set identified. We measured the discrepancies using accuracy (see Table 5). In the table, accuracy reports the number of correct identifications of topics in the gold-standard set documents compared to the total number of identifications (expressed as a percentage). We considered the topic modeling classification correct if it matched the one that humans performed.

4.2 Assessing the Different Models against the Gold-standard Sets

Figure 3 represents the 150 models' accuracy in comparison to the three gold-standard sets. On the Y-axis, we report the models' accuracy (as a percentage), while, on the X axis, we plot the alpha values that the topic model used. The numbers in the chart indicate the different delta values. The models' accuracy decreased noticeably as the alpha values increased. This result remained consistent regardless of the gold-standard set we used. We do not find this result surprising because smaller alpha values allow the topic modeling algorithm to concentrate the topic distributions only on a single (or a few) topic for each document. Such freedom allows the algorithm to identify the most dominant topic for each sentence in the review or response.

In Table 5, we report the best and worst performing model parametrizations for each gold-standard set.

The table shows that Model 63 performed the most accurately for both the inclusive and the full agreement set (specifically, its accuracy ranged from between 78.3% and 80.4%). Model 136 performed the most accurately (76.8%) for the partial agreement set. While Model 63 did not perform the best for this set (76.0%), it still fell among the top ten most accurate models. When it comes to the worst model for all three sets, Model 90 performed the least accurately. This model had an accuracy value between 4.4 and 5.6 percent smaller than the best model. Most importantly, Table 5 shows we can easily compare models with different parametrizations by using the gold-standard set and identify the model that performs the closest to human labeling—the optimal model in relation to the topics' human interpretability.

Researchers have to make a decision on which gold-standard sets to use to determine the optimal model based on what assessment they want to perform on the topic modeling results. The inclusive set, for example, exposes the algorithm to the same difficulty that humans encounter in labeling the different documents. When researchers compare the results from different topic models against the inclusive set, they can benchmark their models against a realistic assessment of how the algorithm performs in a task that humans completed. We recommend using the inclusive set when researchers want to benchmark different models' accuracy against realistic human-classification tasks in organizations (e.g., hotel managers attempting to gauge the service level in their firms). However, if researchers want to assess the topic modeling results' interpretability against a set of documents and/or topics that lack disagreement (i.e., that human raters unanimously recognized), we suggest using the full or the partial agreement sets. In fact, by comparing different models' results against the full agreement set, researchers can benchmark how their algorithms perform against a set of documents for which validators identified topics in a consistent way (i.e., "clear cut" documents). Generally, researchers should expect the models to classify the full agreement set with a higher accuracy than the other sets because it includes topics and documents that humans unequivocally identified. We recommend researchers use the full agreement set when they focus on assessing model performance in a task in which humans reached full consensus. Finally, by using the partial agreement set, researchers can assess how the models perform by classifying

only those topics that received full consensus among human raters. We recommend the partial agreement set for benchmarking tasks in which researcher want to make sure the algorithm can identify only those topics that a set of documents clearly contain.

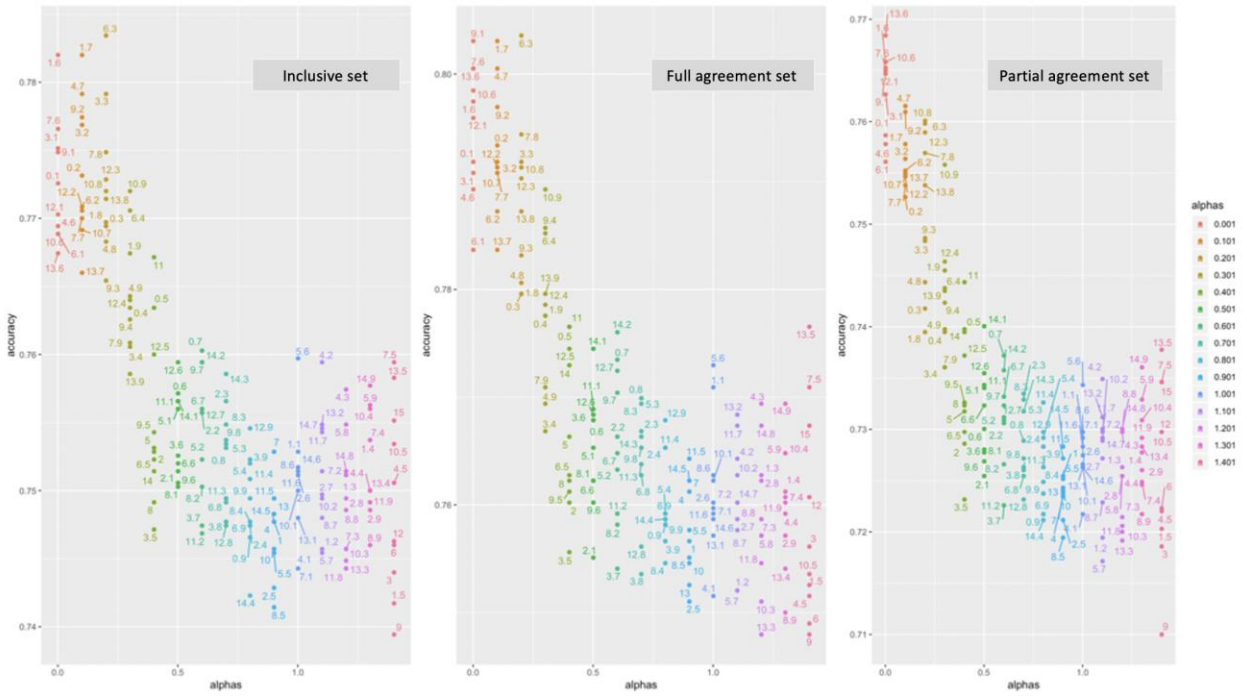

Figure 3. Topic Models' Accuracy Comparisons

Table 5. Models with the Highest and Lowest Accuracy

Models	Inclusive set			Full agreement set			Partial agreement set		
	Accuracy	Alpha	Delta	Accuracy	Alpha	Delta	Accuracy	Alpha	Delta
Best model	78.3%	0.201	6.3	80.4%	0.201	6.3	76.8%	0.001	13.6
	#63			#63			#136		
Worst model	73.9%	1.40	9	74.8%	1.40	9	71.4%	1.40	9
	#90			#90			#90		

We believe that researchers should also assess the models against all three sets so they can think about the performance of the different models' parametrizations. Moreover, it can help researchers understand which parametrization is more efficient in which situation. Lastly, if a model consistently outperforms the others, that justifies the decision to select it regardless of one's research objective. In fact, in the above illustration, researchers should select Model 63 for future analysis because it performed the best regardless of the gold-standard set that we used for benchmarking. In other words, Model 63 can replace human classification with an accuracy of 78.3 percent in contexts in which researchers emphasize inclusiveness among topics in a document collection (inclusive set), with an accuracy of 80.4 percent in contexts in which researchers emphasize full agreement among all topics in a document collection (full agreement set), and with an accuracy of 76.0 percent in contexts in which researchers emphasize agreement among certain topics in a document collection (partial agreement set). HIT can not only extract human-interpretable topics but also confidently scale the task to large corpora−corpora that humans could not feasibly tag. Our best performing model labeled all 87,962 documents available in our corpus in minutes.

5 Guidelines and Recommendations

HIT represents an innovative method for selecting models when researchers want to assess identified topic's human interpretability. HIT ensures the topics that researchers algorithmically extract from an entire corpus match the topics that human raters would have identified in the same corpus with high probability. Our approach demonstrates that one can rigorously identify optimal model parametrizations for maximum interpretability and rigorously justify model selection. HIT provides validity and trustworthiness to topic modeling results. In this section, we offer several guidelines and suggestions to help researchers use and adapt our approach in the future.

5.1.1 Know Your Topics before You Start Any Analysis

HIT requires that researchers know a priori which topics pertain to their research. Thus, they can use it for weakly supervised and supervised topic modeling but not for unsupervised topic modeling. Researchers should invest time in determining what topics pertain to their study context. We suggest that, to select topics, researchers should review the relevant literature, conduct a preliminary analysis, or do both.

5.1.2 Choose Appropriate Seed Words

Seed words play a critical role in extracting topics of interest. For this reason, researchers need to pay close attention to selecting them. We recommend that researchers preliminarily analyze the most frequent terms in the corpus they analyze and then identify those terms that orthogonally represent each topic relative to the others. In this step, researchers should focus on choosing those words that maximally discriminate topics. Finally, in general, researchers should select only nouns as seed words because adjectives and adverbs generally appear across topics.

5.1.3 Look for Existing Gold-standard Sets

Researchers require gold-standard sets to adopt HIT. As we show above, gold-standard sets enable one to systematically and rigorously evaluate different topic models' parametrizations and to measure their precision in identifying human-interpretable topics. Moreover, the human validated gold-standard sets should reflect the context that the researcher has an interest in. We believe the *AIS Transactions on Replication Research* represents the perfect outlet to publish and collect open access gold-standard sets. We hope researchers will follow our example and make publicly available gold-standard sets in different areas of interest. However, when they cannot access context-specific gold-standard sets, researchers can replicate the procedure that we discuss in this paper to create them.

5.1.4 Be Careful about Labeling New Gold-standard Sets

Labeling the documents in the gold-standard set involves some complexity because classifying documents according to the topics they contain requires researchers to extensively train raters. Other researchers who want to validate their topic-extraction approaches according to human interpretability can adopt and adapt the procedures we followed. Keep in mind that, in case the inter-rater agreement after each step does not reach the desirable threshold (> 0.61 Fleiss' kappa), researchers might need to hold extra consolidation meetings. Moreover, given that the classification process can consume much time and money, we encourage researchers to make such labeled sets available to help other researchers use this method in future studies. Once researchers produce gold-standard sets, others can rapidly generalize and scale model validation.

5.1.5 Use the Gold-standard Sets to Benchmark Models and Not to Train Them

Gold-standard sets usually represent a small subset of a corpus. Thus, generally speaking, one cannot use them to train models. However, researchers can use them as a holdout benchmarking set to which they assess the human interpretability of results from models with different hyperparametrizations. This model comparison against human labeling not only enables researchers to identify and justify model parametrization but also demonstrate the model's accuracy in completing a task unfeasible for humans. To avoid overfitting, researchers should not include gold-standard sets in the corpus they use to run the LDA algorithm. In fact, researchers will use the posterior of the topics' proportions to label unseen documents (those in the gold-standard sets). By doing so, researchers will have a version of the gold-standard set that the algorithm labeled and one that humans labeled. Comparing the two allows one to

assess different model parametrizations' accuracy. HIT combines human coding's advantages with regard to topic interpretability with topic modeling's analytical efficiency and scalability.

5.1.6 Do not Limit HIT to LDA

We believe the approach that we propose here generalizes to other types of documents and to other text-mining techniques. Researchers can also use gold-standard sets to evaluate a text-mining technique's ability to extract topics from other types of textual documents (e.g., transcripts of service bots). Moreover, we envision that researchers could apply gold-standard sets to compare and benchmark competing techniques beyond LDA, such as support vector machines and latent semantic indexing.

6 Conclusions

In this paper, we propose a new method to address criticism that researchers benchmark different topic models and determine results' human interpretability haphazardly rather than systematically (Eickhoff & Neuss, 2017). HIT offers a scalable approach to evaluate the degree to which humans can interpret extracted topics in a corpus, which increases topic modeling results' validity and trustworthiness. We illustrate how to evaluate different topic models' performance using HIT and demonstrate how, even on a large corpus, the topic modeling results meaningfully represent the topics that the unstructured text data discuss. Our approach ensures the topics that one extracts algorithmically from a corpus align with the topics human raters would have extracted from the same corpus. We also contribute a protocol for creating reliable gold-standard sets that researchers can use to benchmark different topic models. Finally, we contribute three gold-standard sets that researchers can use to analyze online customers' reviews in the hospitality context.

References

Abbasi, A., Hassan, A., & Dhar, M. (2014). Benchmarking twitter sentiment analysis tools. In *Proceedings of the Language Resources and Evaluation Conference*.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17-35.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). *Care and feeding of topic models: Problems, diagnostics, and improvements*. Boca Raton, FL: CRC Press.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the Neural Information Processing Systems Conference*.

Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems*, *39*, 110-135.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ArXiv.* Retrieved from http://arxiv.org/abs/1702.08608

Eickhoff, M., & Neuss, N. (2017). Topic modelling methodology: Its use in information systems and other managerial disciplines. In *Proceedings of the European Conference of Information Systems*.

Fei-Fei Li, & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

Feinerer, I., & Hornik, K. (2018). tm: Text mining package (Version 0.7-6). Retrieved from https://CRAN.R-project.org/package=tm

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(1), 1-54.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(1), 1-30.

Grün, B., & Hornik, K. (2018). topicmodels: Topic models (ver. 0.2-8). Retrieved from https://CRAN.R-project.org/package=topicmodels

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lu, B., Ott, M., Cardie, C., & Tsou, B. (2011). Multi-aspect sentiment analysis with topic models. In *Proceeding of the 11th International Conference on Data Mining Workshops.*

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Mullen, L., Keyes, O., Selivanov, D., Arnold, J., & Benoit, K. (2018). tokenizers: Fast, consistent tokenization of natural language text (ver. 0.2.1). Retrieved from https://CRAN.R-project.org/package=tokenizers

Müller, O., Junglas, I., Brocke, J. vom, & Debortoli, S. (2016). Utilizing big data analytics for information systems research: Challenges, promises and guidelines. *European Journal of Information Systems*, *25*(4), 289-302.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Proceedings of the Human Language Technologies Conference*.

Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010). Evaluating topic models for digital libraries. In *Proceedings of the 10th Annual Joint Conference on Digital libraries.*

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88-102.

Palese, B., & Usai, A. (2018). The relative importance of service quality dimensions in e-commerce experiences. *International Journal of Information Management*, *40*, 132-140.

Piccoli, G. (2016). Triggered essential reviewing: The effect of technology affordances on service experience evaluations. *European Journal of Information Systems*, *25*(6), 477-492.

Piccoli, G., and Ott, M. (2014). Impact of mobility and timing on user-generated content. *MIS Quarterly Executive*, *13*(3), 147-157.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and Costs. *American Journal of Political Science*, *54*(1), 209-228.

TripAdvisor. (2017). *Q4 2016 results.* Retrieved from http://ir.tripadvisor.com/static-files/d555b056-765f-463d-b27b-a6d3cfcf5ad4

TripAdvisor. (2019). *Q4 2018 results*. Retrieved from http://ir.tripadvisor.com/static-files/6d4c71fd-3310-48c4-b4c5-d5ec04e69d5d

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *In Proceedings of the 26th Annual International Conference on Machine Learning*.

Appendix A: A Classification Procedure for Gold-standard Sets

This classification procedure comprises two distinct parts. In the first part, we describe and provide examples of seven different topics. In the second part, you will need to use them to classify a list of documents (available in Excel). You have completed the classification when you have assigned the topics to all the documents in the list. Keep in mind that each document can contain a different number of topics. You need to report all the topics that each document discusses according to your judgment.

Part 1

To complete this procedure, you need to understand each topic. Please, read the descriptions and examples below carefully before you begin classifying the documents available in the Excel file.

Service: comments that describe the service that the hotel provides. It includes comments that describe the hotel staff's performance in helping guests with any element of their stay (e.g., check-in, check-out, problem resolution, etc.).

Examples:

1) "The service at this hotel is very good, and they are very polite and friendly."
2) "We will strengthen the personnel training to provide the impressive service for all guests."
3) "Check-in is quick and the staff friendly."

Value: comments that refer to the hotel's economic value proposition. It includes comments that describe how guests perceived the experience they received for the price they paid. It also contains general assessments about the stay without including the price.

Examples:

1) "The price-value is excellent."
2) "I think this hotel is fantastic value for money and would recommend to anyone."
3) "Totally worth it for not much more money."
4) "Overall, everything was good."

Location: comments that describe the hotel's location and its surroundings, such as the view from the hotel or local attractions.

Examples:

1) "The location is great as it is close to two subway stations and is located in the nice area of the city."
2) "Close to restaurants and bars."
3) "Huge shopping malls are within walking distant."

Room: comments that relate to the room's physical aspects, amenities, and areas (e.g., the bathroom).

Examples:

1) "The rooms are spacious but noisy."
2) "LED TV and a comfortable bathroom."
3) "The bed is custom made by Simmons and is very comfortable."

Food: Comments that describe food or drinks served in the hotel and its restaurants, which includes quality of the breakfast buffet or the food/drink delivered to the room.

Examples:

1) "They only had 2 kinds of dressings for salad."
2) "The food was surprisingly delicious for breakfast and dinner."
3) "The Chinese restaurant in the hotel is also wonderful."

Greetings and salutations: comments that contain polite words or signs of welcome or recognition. Hotel responses often include them, but they can also appear in reviews (e.g., Dear GM of the Royal Hotel…).

Examples:

1) "Best regards, General Manager"
2) "Thank you for your continued support.
3) "With warm regards, Hotel."

Hotel amenities: comments that relate to hotel facilities not in the room, such as the gym, the pool, the spa, etc.

Examples:

1) "Pool: There is an outdoor pool on the 5th floor with ropes to swim laps."
2) "Gym: This is one of the best hotel gyms ever."
3) "There is a yoga/pilates/spin studio."
4) "WIFI is not free"…

Unclassifiable: comments that you cannot confidently put in any other category.

Examples:

1) "I was very happy!"
2) "As usual, staying at here is like coming to a 2nd home."
3) "All the wonderful things I've read in this forum are all true."

Part 2

Keep in mind the above descriptions and examples while completing the classification in the Excel file "goldSetRevRes". If you think a comment contains more than one topic, report all them in order of dominance. However, remember to insert only one topic per column. For example, we have a comment that includes four different topics: topic 1 = value; topic 2 = service; topic 3 = location; topic 4 = unclassifiable.

When you have classified all the documents, please save the file as "yourfullname_goldSetRevRes" (e.g. joshadms_goldSetRevRes) and email it to the lead researcher. You have now completed the classification. Thank you.

Appendix B: Topic Models Seed Words

Table 6. Seed Words that We Used to Fit the Topic Models

Value	Service	Room	Location	Hotel amenities	Greetings and salutations	Food
Stay	Staff	Bed	Place	Gym	Regards	Breakfast
Experience	Reservation	Bathroom	Area	Spa	Guest	Restaurant
Value	Service	Room	Location	Pool	Dear	Food
Recommend	Desk	Shower	View	WiFi	Forward	Bar
Price	Help	Spacious	Distance	Parking	Feedback	Buffet
Money	Unresponsive	Interior	Located	Internet	Comments	Dining
	Refuse	Housekeeping	Convenient	Club	Sincerely	Eggs

About the Authors

Biagio Palese is an Assistant Professor for Information Systems in the College of Business at Northern Illinois University. He earned his PhD in Business Administration with a concentration in Information Systems at Louisiana State University. His teaching and research interests embrace introduction to management information systems, data analytics, effective use, customer service, digital data streams and text mining. His research has appeared in journals such as *MISQ Executive*, the *International Journal of Information Management*, the *European Journal of Information Systems, Information & Management* and in conference proceedings, including the International Conference of Information Systems and Americas Conference of Information Systems. The potential of his research has been recognized with his selection at the ICIS 2018 Doctoral Consortium. He has presented at various conferences, including ICIS, AMCIS, SIM Connect Live and BIG XII MIS Symposium.

Gabriele Piccoli is the Edward G. Schleider chair for information systems in the E. J. Ourso College of Business at Louisiana State University and is on leave from the University of Pavia. He is the director of the Digital Data Streams Lab at LSU. His research, teaching, and consulting expertise are in strategic information systems and the use of advanced IT to support customer service. His most recent research focus is on digital data streams and their potential for value creation. He is author of the book *Information Systems for Managers: Text and Cases*. His research has appeared in both academic and applied outlets such as *MIS Quarterly, Journal of AIS, European Journal of Information Systems, Decision Sciences, California Management Review, MIS Quarterly Executive, Communications of the ACM*, and *Harvard Business Review*. He has published 16 full-length teaching case studies through Communications of *the Association for Information Systems* as well as *Harvard Business School Publishing*.