World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:8, 2016

Mining User-Generated Contents to Detect Service Failures with Topic Model

Kyung Bae Park, Sung Ho Ha

***Abstract*—**Online user-generated contents (UGC) significantly change the way customers behave (e.g., shop, travel), and a pressing need to handle the overwhelmingly plethora amount of various UGC is one of the paramount issues for management. However, a current approach (e.g., sentiment analysis) is often ineffective for leveraging textual information to detect the problems or issues that a certain management suffers from. In this paper, we employ text mining of Latent Dirichlet Allocation (LDA) on a popular online review site dedicated to complaint from users. We find that the employed LDA efficiently detects customer complaints, and a further inspection with the visualization technique is effective to categorize the problems or issues. As such, management can identify the issues at stake and prioritize them accordingly in a timely manner given the limited amount of resources. The findings provide managerial insights into how analytics on social media can help maintain and improve their reputation management. Our interdisciplinary approach also highlights several insights by applying machine learning techniques in marketing research domain. On a broader technical note, this paper illustrates the details of how to implement LDA in R program from a beginning (data collection in R) to an end (LDA analysis in R) since the instruction is still largely undocumented. In this regard, it will help lower the boundary for interdisciplinary researcher to conduct related research.

***Keywords*—**Latent Dirichlet allocation, R program, text mining, topic model, user generated contents, visualization.

I. Introduction

IT is no longer a surprise that online UGCs impact our daily lives in many ways, especially for shopping and traveling. Consumers are benefiting from obtaining the views of similar other consumers prior to making purchasing decision. Contrary to consumers, it opens up a new challenge for the management to handle the data tsunami expressed online from consumers.

If management can identify their defects of products or services in general, or customer complaints at a scale in specific, then management can categorize and prioritize them in a timely manner considering negative news travel faster than positive ones [1]. Eventually their brand image could be efficiently under control, increasing the effectiveness to manage their weaknesses given limited capacity and resources. Unfortunately, a current approach (e.g., sentiment analysis) is sometimes limited and ineffective for leveraging textual information to detect the defects or problems at stake that a certain management suffers from. To tackle such issue, we demonstrate one way of how to answer such question by employing LDA, one of the most successful topic models, to detect service failures. The key idea is that the sorted customer complaints through topic model reveal the indication of the service failures, prompting management's effort to recover the failures.

To contribute for a broader technical note, this paper also provides and illustrates the whole procedure to implement and conduct LDA analysis from data collection on the web to final analysis together with visualization technique, especially in R Program. In this regard, this adds incremental knowledge in that now many documentations are still available of implementing LDA analysis in R program. Hopefully, it will help open the door for the entry interdisciplinary researcher relatively easy other than machine learning community.

II. Literature Review

Review Mining Studies

Early studies [2] found strong relationship between reviews and firm sales. If customers continue to mention about certain products (or services), more customers become aware of those products. Moreover, if customers score positive than negative on the products, it helps induce buying decision from prospective customers. Stated differently, the main observations from earlier studies that the valence and number of reviews impact positively on firm sales [3] elicited extant literature on review mining. Toward the same direction, a parallel stream of research begins to focus on the review text, often ignored due mainly to methodological difficulty but major component of review in earlier studies, and incorporating reviews for further analysis has attracted many researchers since it provides near-genuine information for a question of what others think.[4] In particular, [3] argues that the valence and number of reviews expressed as numerical numbers cannot fully capture the natural heterogeneity of each customer, and posit the importance of incorporating the text itself for analysis to be much more meaningful. As if reflecting such argument, tangentially related voluminous research on methodological development has progressed, namely opinion mining or sentiment analysis in text mining. Broadly speaking, it classifies the state of a polarity for emotion or opinion of people into positive or negative, and has been further developed to detect the polarity based on product feature (so called, aspect-level) rather than review as a whole (document-level).

With successful development in terms of methodology, extant studies have employed sentiment analysis to diverse domains such as product reviews, movie reviews, political

KyungBae Park works at Management Information System, Kyungpook National University, Daegu, South Korea (phone: +82-10-7213-1871; e-mail: iamkbpark@knu.ac.kr).

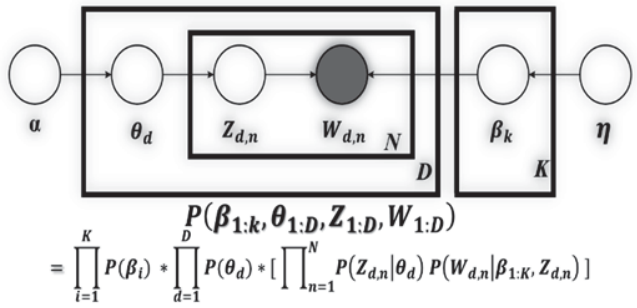
Sung Ho Ha is a professor of Management Information System at Kyungpook National University, Daegu, South Korea (phone: +82-53-950-5440; e-mail: hsh@knu.ac.kr).

orientation extraction, and stock market predictions. To date, sentiment analysis in general has been extended from document-level, sentence-level, to feature-level in the order of methodological complexity with better capability in analysis [4].

LDA

Meanwhile, a generative model approach on text mining introduces LDA as one of the simplistic form of probabilistic topic models to detect valuable patterns or topics in document corpora [5]. LDA intuitively assumes that documents, a mixture of corpus-wide topics, exhibit multiple topics. Here, topics are defined as a distribution or matrix of words, and words are drawn from one of those topics.

The application of LDA has been widespread in that it automatically discovers the topics from a large collection of electronic archives. Within the widely researched sphere since the introduction of LDA, a plethora of research has been developed to better understand the opinions expressed in textual documents with the methodological improvements. As a probabilistic graphical model, the generative process of LDA is the joint distribution and is expressed as follows (see Fig. 1):



$$P(\beta_{1:k}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{i=1}^{K} P(\beta_i) * \prod_{d=1}^{D} P(\theta_d) * \left[\prod_{n=1}^{N} P(Z_{d,n}|\theta_d)\, P(W_{d,n}|\beta_{1:K}, Z_{d,n}) \right]$$

Fig. 1 A graphical model of LDA: K = Number of Topics; D = Number of Documents; N = Number of Words; $\beta_{1:k} = Topics, where\ each\ \beta_k\ is\ a\ distribution\ over\ the\ vocabulary$; $\theta_d = Topic\ Proportions\ for\ the\ d^{th}\ document,$ $where\ \theta_{d,k}\ is\ the\ topic\ proportion\ for\ topic\ k\ in\ document\ d$; $Z_d = Topic\ Assignments\ for\ the\ d^{th}\ document,$ $where\ Z_{d,n}\ is\ the\ topic\ assignment\ for\ the\ n^{th}\ word\ in\ document\ d W_d = Observed\ Words\ for\ the\ document\ d,$ $where\ W_{d,n}\ is\ the\ n^{th}\ word\ in\ document\ d,$ $which\ is\ an\ element\ from\ the\ fixed\ vocabulary$

In reality, we only observe words in document, and we therefore infer hidden variables of topics, proportions, and assignments given documents. From the perspective of grasping the clusters of management deficiencies from review text, LDA meets our research purpose. Hence, we proceed to discussion of fundamental procedures to conduct the analysis, and implement LDA in R, one of the popular software, followed by the demonstration of analysis results.

III. METHOD

Implementation LDA in R Program

We use R, freely available open source statistic and machine learning software because it has recently gained burgeoning popularity in academia and practitioner. Though there is a considerable body of literature on LDA, limited attention has been devoted to show the actual procedures to conduct the analysis in R. The manuals are still lacking except the handful of documents [6], to the best of our knowledge. It replicated and implemented LDA analysis mainly in R for the renowned previous paper on LDA, [7] introducing Gibbs sampling inference technique unlike variational methods used in original LDA paper. [5] Even in the existing document, however, additional software, *Python*, was used to handle the textual data which takes a great part prior to the ultimate analysis. Paradoxically, these shortcomings of a technical note in existing research are problematic for novice researchers who would like to employ LDA but lacks technical knowledge. In response to the shortcomings, our paper provides the comprehensive guide as additional contribution with compact illustration. In reality, apart from the general coding guidelines for the functions provided in each package manual, the explanations are seldom stated to correct errors from step to step. This is problematic and creates steeper learning curve, thereby technically frustrating researchers. As with other analysis, the difficulty is not to apply the codes in general, but to correct the errors in specific. Therefore, we systemically created piecewise codes for each implementation steps to handle generally observed errors as much detail as possible.

Relatedly, there are several tools available to implement LDA such as Mallet in Java, Gensim in Python, Mahout in Hadoop, and recently added library in Spark. However, in order to lower the barrier for the entry researcher other than machine learning community, we intentionally implement the LDA in R from data collection to final analysis. The application of LDA has paved the way for wide coverage in social science (i.e. communication or marketing).

Since suggesting to use the combination of several softwares may overwhelm them to begin with, providing the implementation under one platform also enables the analysis succinctly. To that end, the current paper describes how to implement LDA analysis from the beginning to the end: data collection, pre-processing, model fitting, and visualization within R program.

Data Collection

Though several tools are available, R program is powerful enough to extract textual data (i.e. user review on products and services) from the web. The packages necessary to scrape the textual data are 'xml2', 'plyr', and 'rvest'. Among them, 'rvest' package takes a central role in that it simplifies the coding scheme to extract review data along with SectorGadget which can easily identify the proper location of each contents to extract.

Pre-Processing

After the extraction of necessary information and prior to employing the appropriate text mining analysis, pre-processing is one of the most significant steps though the descriptions of how are often overlooked in many documentations. 'tm' is the package enabling such task in R. In LDA analysis, following

techniques are applied: (1) deleting words whose number of characters are less than two and greater than forty-five, (2) removing numbers, punctuations, stop-words, and whitespace, and (3) transforming upper case letter to lower.

In fitting LDA model after pre-processing, the input requires the format of document-term matrix (DTM) with term-frequency (TF). In addition, terms appeared at least five documents (i.e. reviews) are considered for analysis. [7] In summary, extracted reviews are converted to a corpus to apply the pre-processing techniques. Then, the final DTM with TF for fitting LDA is created.

Model Fitting

Two major packages are available to employ LDA in R: 'lda' and 'topicmodels'. Although these two are commonly used and performance is comparable, 'topicmodels' is preferable in that the documentation (package manual) is readily available rather than 'lda'. To fit the model, there are two important sub-procedures to be taken: (1) finding the best number of topics (k), and (2) fitting LDA model again with the found best number of topics.

First, to find the optimal k, models are fit from the iteration of two to a hundred twenty topics. Even though selecting the numbers of topics to run depends on the volume of texts with several hundreds of topics shown on literature, limiting the iterations to 120 topics is applied to find the practical use case of interpreting topics. Commonly used Gibbs sampling inference technique is applied since it is unbiased and relatively easy to understand. One caveat is that Markov chain Monte Carlo is sometimes computationally inefficient when working with large corpora. [8] From the repetitions of fitting the models, resulting log-likelihood values for each number of topics are obtained and compared to discover the highest value: the number of topics with highest log-likelihood value indicates the optimal number of k, and 'ggplot' package is used to graphically spot the k.

Second, the final LDA is re-fit with found optimal k. To note, hyper-parameters in LDA such as α and β are set as follows as suggested [7]: $\alpha = 50 / k$, and $\beta = 0.1$. Those are the default values of LDA function in 'topicmodels' and no specification is necessary.

Visualization

One of the major difficulties in LDA is to practically interpret the topics found for the analysis. To overcome such obstacle, several visualization techniques have been proposed (i.e. Termite). Recently, 'LDAvis' package [9] in R was introduced to help analysts better understand the topics with interactive graphical representation. The major distinctive feature in 'LDAvis' package is λ-adjustable value setting, considering the relative importance of topic-specific term. Unlike other visualization techniques, 'LDAvis' intuitively displays resulting topics with keywords to analysts with relevancy measure (λ) to seamlessly interpret the topics. The outputs of LDA function of 'topicmodels' package used to fit model can be nicely fed into 'LDAvis' as input parameters.

Naming Topics for Interpretation

Finally, naming topics is the last step though it could be optional. To obtain the reasonable names of each topic, cross-check with other researchers is often helpful. Clearly, the consult from domain experts may strengthen the soundness of topic name.

Here, we use Amazon Mechanical Turk (AMT), online labor market, to cross check with topics found in order to properly name the topic. We ask three people for each topic to name and eventual names for each topic were then aggregated and finally derived by researcher.

Analysis

To demonstrate the actual implementation of LDA in R program environment, this section provides analysis. We employ LDA on reviews of customer complaints for one of the major logistics companies in the USA from consummeraffairs.com in order to detect service failures as candidate topics for analysis.

We first collected user-generated reviews. Here, reviews by consumers are mainly dedicated to complaints. As such, it is reasonable to assume that the resulting topics may signal the potential service defects. The entire review data available were crawled on Feb 23, 2016, and the total number of reviews were 6,083 at the time of extraction, and 4,997 words were returned after pre-processing. Each review was treated as a single document in the corpus, and was pre-processed to create DTM.

Next, to find the optimal number of topics, we applied $\alpha = 50 / k$ and $\beta = 0.1$ for all runs of algorithm. Then, plotting them to spot the highest log-likelihood value reveals the topics with 109, as shown in Fig. 2. However, starting from the 70 topics, the values become flat with no major difference in log-likelihood values, and optimal number of topics used is 70. In summary, the fitted LDA model for the corpus returned after Gibbs sampling (method = 'Gibbs') of 2,000 iterations (iter = 2,000), and hyper-parameters of $\alpha = 50 / k$ (documents) and $\beta = 0.1$ (topics), and number of topics (k=70).

Upon fitting the final LDA model, there are several outputs (slots) as a result including log-likelihood value. Among them, the function to implement visualization in 'LDAvis' package requires the inputs of *phi, theta, vocab, doc.length*, and *term.frequency* with *R* as an optional parameter: *R* was set as 20 to display top 20 keywords for each topic. Fig. 3 shows the sample screenshot of topic number 1 in 70 topics.

With an adjustment of λ value (relevancy metric) from 0 to 1, the bar graphs of blue-bar and red-bar allow analysts for the easier navigation of keywords, and in turn, the interpretation of topics. Blue-bar shows the overall term occurrences across the entire 70 topics, often not very helpful to understand the topic. Yet, red-bar reveals the terms' higher relevancy for the specific topic, triggering the cues to interpret the topic. In this way, analysts can easily identify dominating keywords for interpreting the topics. Due to space limitation, the table with all 70 topics and 20 keywords are omitted. Instead, the example table with 10 topics is shown in Table I.

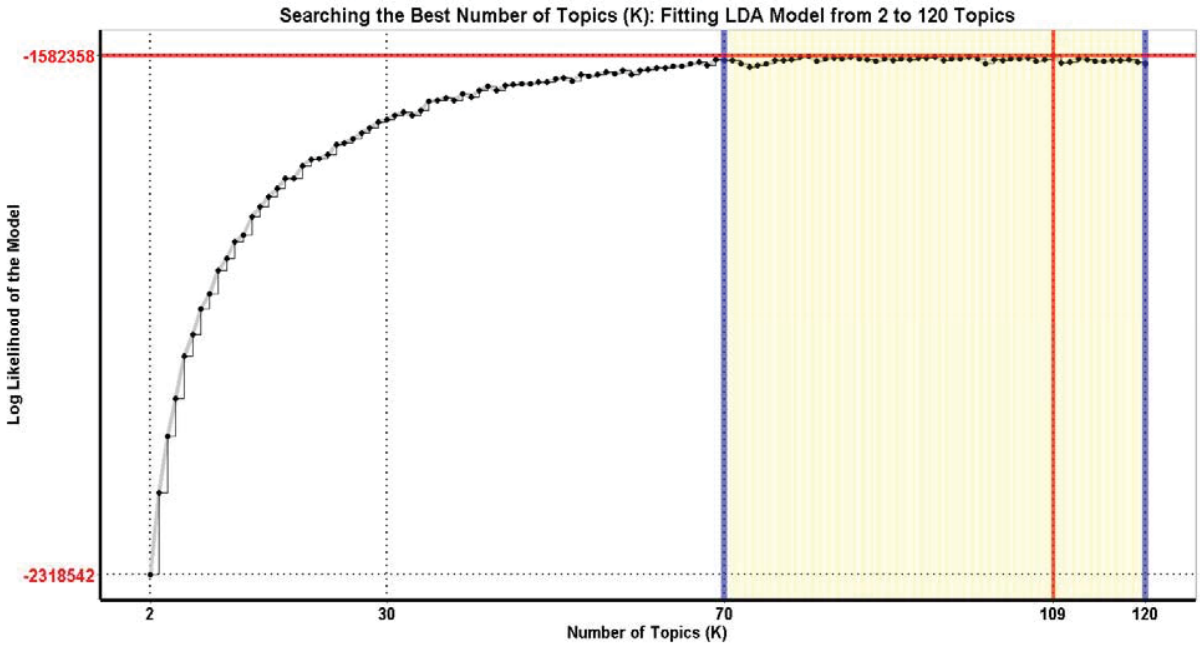


Fig. 2 LDA model fitting from 2 topics to 120 topics (ggplot)

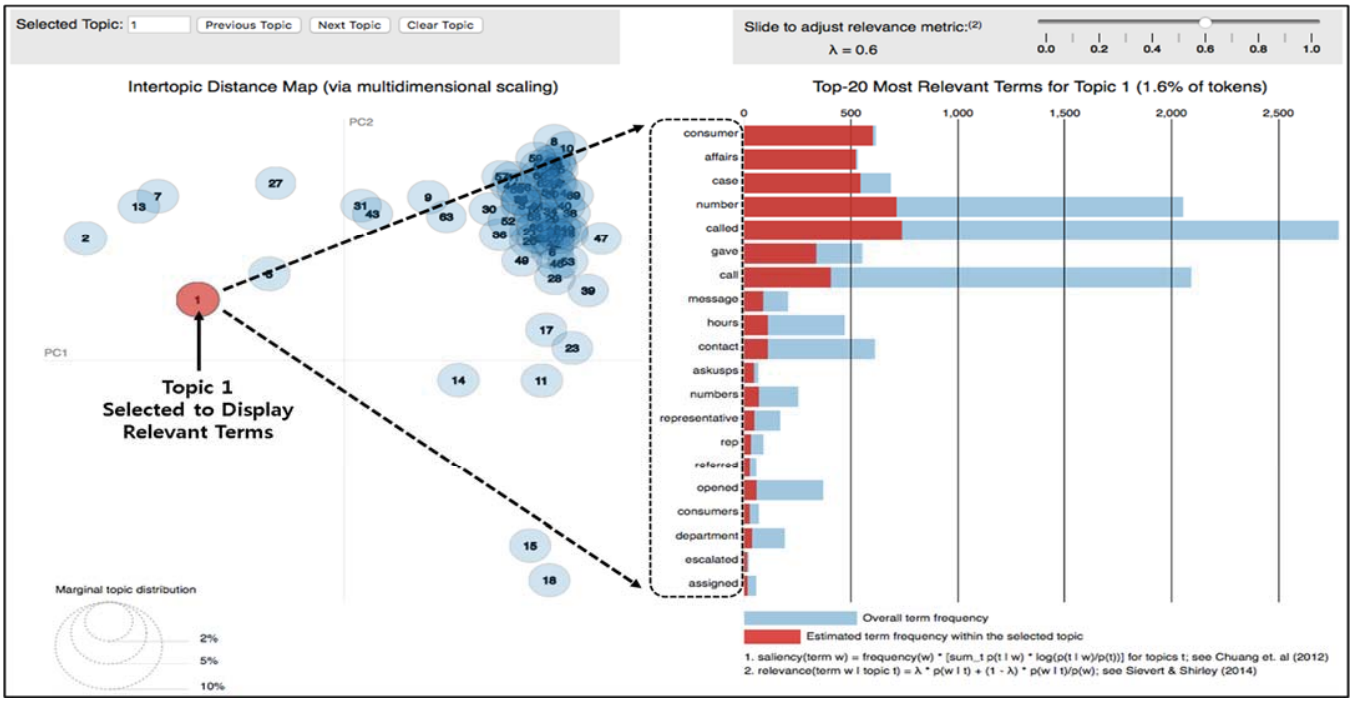


Fig. 3 LDAvis output. Here, topic 1 (red circle on the left) was selected to display 20 relevant terms, defining topic 1

TABLE I
AN EXAMPLE OF TOPICS AND KEYWORDS WITH THE ODER OF IMPORTANCE

Topic 3	Topic 17	Topic 21	Topic 30	Topic 38	Topic 46	Topic 51	Topic 52	Topic 55	Topic 68
Address	Item	Customer	Returned	Late	Items	Money	Form	Pay	Card
Change	Refund	Service	Sender	Bills	Lost	Order	Center	Paid	Credit
Forwarded	Ebay	Poor	Address	Bill	Missing	Lot	Passport	Cost	Bank
Moved	Shipped	Provide	Return	Payment	Stolen	Loss	Fill	Extra	Account
Forwarding	Seller	Worst	Addressed	Due	Worth	Lost	Atlanta	Fee	Statements
Changed	Shipping	Horrible	Correct	Fees	Responsible	Losing	Filled	Amount	Statement
Forward	Buyer	Rep	Undeliverable	Payments	Stealing	Orders	Application	Charge	Theft
Previous	Product	Expect	Marked	Mortgage	Stole	Spend	Needed	Additional	Cards
Submitted	Bought	Representative	Error	Charges	Responsibility	Replace	Recovery	Price	Charged
Family	Sold	Experienced	Returning	Companies	Steal	Food	Processing	Fault	Identity
Forms	Shipment	Skills	Physical	Company	Theft	Spent	Process	Rates	Green
Form	Purchased	Terrible	Senders	Junk	Merchandise	Afford	Appointment	Paying	Cash
Temporary	Paypal	Disappointed	Correctly	Stop	Investigate	Gas	Mailing	Taxes	uscis
Move	Shipper	Everyday	Addressee	Pay	Insure	Lose	Distribution	Dollars	Debit
Sticker	Items	Agents	Incorrect	Interest	Finding	Dress	Visa	Send	Fraud
Moving	Ordered	Ways	Verified	Water	Sentimental	Refunded	Camera	Services	Recently
Completed	Buyers	Helpful	Back	Utility	Contained	Result	Search	Include	Worry
Yellow	Sell	Useless	Unknown	Paying	Coins	Funds	Trip	Mention	Including
Filled	Total	Ohio	Insufficient	Checks	Thieves	Paying	Section	Higher	Billing
Receiving	Ship	Bucks	Deliverable	Causing	Risk	Sick	Renewal	Pocket	Requested

Note: For representation purpose, pre-processed lower texts were converted back to regular word form (i.e. 'Address' was used instead of 'address' in table)

IV. RESULT

70 topics drawn from LDA might be less appealing to managers since it seems myriad in a practical manner. Thus, rigorous investigation of each topic was necessary to warrant the findings sounding. It reveals that several meaningless topics were found. Such topics thwart providing valuable interpretation although topic themselves are firmly cohesive in lexicosemantic perspective. For example, topic 25 is nicely derived with the top 20 keywords but are date-related words only such as ‘day’, ‘Monday’, ‘Sunday’, ‘Friday’, ‘October’, and so on. Algorithmically, LDA performed well on what it was supposed to do. Practically, however, such topic turned out to be noisy since it does not provide any meaningful insight. By the same token, several uninterpretable meaningless topics were also found and excluded for the final analysis, and this returned 48 topics.

Consistent with expectations, the most frequent topics found are the ones related to customer service and delivery. This makes intuitive sense in that what logistics company mainly offers customer is a delivery service.

It is found that various customer complaints expressed online were successfully derived into the groups or related topics by the implementation of LDA together with one of the visualization methods, LDAvis, in a more intuitive manner, thereby leading to the better understanding the issues found. In this vein, management can effectively investigate the concerns from their existing and(or) prospective customers more precisely in order to efficiently assign their limited time and resources, and to rectify their service failures to recover.

V. DISCUSSION

Our results tell managers working for product or service firms that various data from UGC with appropriate method can provide opportunity for them though challenged to manage. The preemptive recovery actions developed after identifying service failures in a timely manner help reduce the risk for the negative image being radically exposed and spread to potential customers. Particularly, the research also has implications beyond the remedying customer complaints. In other words, management can tactically use identified complaints in a way to improve customer satisfaction.

Our paper provides insights into how analytics on social media can help remedy and maintain the management proactively. Our interdisciplinary approach also highlights insights by applying machine learning techniques in marketing and service management research domain. To promulgate to wider audience including social science domain, we provide succinct theoretical background of LDA and required instruction to implement in R. Hopefully, it could lower the barrier for the entry researchers so that they can fully take the benefit from both. However, there is an important note to mention. Topic model itself is not a panacea for success without careful examination. As shown in our analysis, over 20 out of 70 topics found to be uninterpretable or meaningless. To eschew such pitfalls, therefore, caution is required to correctly understand the topic thereby grasping maximum valuable insights.

REFERENCES

[1] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi, "Bad news travel fast: A content-based analysis of interestingness on twitter," in *Proceedings of the 3rd International Web Science Conference*, 2011, p. 8.

[2] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research,* vol. 43, pp. 345-354, 2006.

[3] N. Archak, A. Ghose, and P. G. Ipeirotis, "Deriving the pricing power of product features by mining consumer reviews," *Management Science,* vol. 57, pp. 1485-1509, 2011.

[4] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*: Cambridge University Press, 2015.
[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.
[6] M. Ponweiser, "Latent Dirichlet Allocation in R," *Theses, Institute for Statistics and Mathematics* WU Vienna University of Economics and Business, Vienna. 2012
[7] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences,* vol. 101, pp. 5228-5235, 2004.
[8] D. M. Blei, "Probabilistic topic models," *Communications of the ACM,* vol. 55, pp. 77-84, 2012
[9] C. Sievert and K. E. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 2014, pp. 63-70