Representativeness of Latent Dirichlet Allocation Topics Estimated from Data Samples with Application to Common Crawl

Yuheng Du, Alexander Herzog, Andre Luckow, Ramu Nerella, Christopher Gropp and Amy Apon
School of Computing, Clemson University, Email: {yuhengd,aherzog,aluckow,rnerell,cgropp,aapon}@clemson.edu

***Abstract*—Common Crawl is a massive multi-petabyte dataset hosted by Amazon. It contains archived HTML web page data from 2008 to date. Common Crawl has been widely used for text mining purposes. Using data extracted from Common Crawl has several advantages over a direct crawl of web data, among which is removing the likelihood of a user's home IP address becoming blacklisted for accessing a given web site too frequently. However, Common Crawl is a data sample, and so questions arise about the quality of Common Crawl as a representative sample of the original data. We perform systematic tests on the similarity of topics estimated from Common Crawl compared to topics estimated from the full data of online forums. Our target is online discussions from a user forum for automotive enthusiasts, but our research strategy can be applied to other domains and samples to evaluate the representativeness of topic models. We show that topic proportions estimated from Common Crawl are not significantly different than those estimated on the full data. We also show that topics are similar in terms of their word compositions, and not worse than topic similarity estimated under true random sampling, which we simulate through a series of experiments. Our research will be of interest to analysts who wish to use Common Crawl to study topics of interest in user forum data, and analysts applying topic models to other data samples.**

***Keywords*-Common Crawl; topic modeling; online forums; unsupervised machine learning;**

I. Introduction

Social media and online forums provide a wealth of data to inform design, engineering, sales and marketing of consumer products. Increasingly, consumers use a wide variety of online services, e. g. Facebook, Twitter, forums, and blogs, to share information and experiences about products and services. Linking product development, sales and marketing to customer needs is a critical capability. The aim of this paper is to investigate the usage of advanced analytics, in particular, natural language understanding techniques, to detect main themes in online forums. Online discussions can help companies to better understand their customers' needs and to improve their products. In comparison to other social platforms, forums can contain very technical and detailed feedback information from advanced users.

Textual analytics approaches are typically based on bag-of-words, n-grams and/or term frequency-inverse document frequency (TF-IDF) [1] data representations. In most cases, these text representations are used in conjunction with classification models, e. g., sentiment classifiers. However, this approach does have some limitations: bag-of-word and TF-IDF representations do not capture themes within documents or semantic relationships. Further, classification approaches require that the data is labeled, which is time-consuming and expensive. In particular, for investigative and exploratory analytics other approaches are better suited, e. g., the ability to capture discussion themes or topics [2]. Topic models [3] are algorithms that can detect common topics across a corpus of documents. The most well-known algorithm is Latent Dirichlet Allocation (LDA) [4]. Topic modeling provides an unsupervised learning method to analyze the main themes in a collection of documents. It can reject the noise in the data and recover the underlying topics hidden in each document without manual tagging. In this paper, we use LDA topic modeling to analyze online forum discussions.

Despite the valuable information provided by online forums, these also have several characteristics that make them intractable to study directly. For example, forums contain a tremendous amount of historical data. Massive online forums such as Gaia Online [5] contain more than 1 billion posts, with a daily count of 20,000 active users. Crawling through the full data of these forums may take months. Continuous crawling of a forum website can also result in blocking of IP addresses. Besides the large size, the directly crawled data are often noisy in nature. Although a forum has structured formats, such as threads and tags to guide discussions, users tend to go off-topic in a thread and spawn multiple discussion themes. Capturing these representative discussion themes requires complex natural language understanding algorithms ([4], [6]). Performing these algorithms on full forum data is a very time-consuming task. Therefore, analyzing large, noisy and complex forum data needs a more efficient strategy.

Common Crawl [7] is an open repository that contains petabytes of web crawl data covering over nine billion web pages [8]. For efficiency purposes, it does not provide the full data of the webpages being crawled. Instead, it provides samples of the online forums in the form of static snapshots. Common Crawl currently performs a monthly crawl based on a two tier crawling strategy which insures that pages with higher page ranks are visited and the overlap between each crawl is minimized. A crawl takes a snapshot of the pages being visited and saves the crawled data into a structured format.

Common Crawl provides a sample of the original online forum data with unknown biases. It is not an independent dataset with respect to the original forum data we are studying. Using Common Crawl as a sample of the full forum data for topic modeling has several advantages: the data is public accessible and ready to use. It avoids many pitfalls that are involved in creating a custom crawler (e. g., the prioritization of web page, blacklisting of the crawler, etc.). Common Crawl snapshots are static, which means they provide consistent data when an analysis requires repeatability. However, one drawback of Common Crawl is the uncertainty with respect to data quality and completeness and thus the ability of using these data for topic modeling. On average, we observed that Common Crawl only contained about 22 % of the data of interest. This is in line with other investigations of the Common Crawl dataset, e. g., by Stolz and Hepp [9]. As the precise collection algorithm of Common Crawl is not known, the data cannot be assumed a true random sample as it may be subject to sampling bias. Thus, it is also not possible to define for which class of data analysis algorithms the data is appropriate and how the results generalize.

In this paper, we evaluate the use of Common Crawl data as a sample for extracting representative LDA topics from online forum textual data. Our focus is on a very active car owner forum that is organized into 14 subforums representing different car models. For each subforum, we collected all available data from Common Crawl – a total of about 280 GB of raw data files. In addition, we developed a customized web crawler to collect the full 2.16 TB data from the online forum. Having access to the full data for each subforum allows us to systematically evaluate the representativeness of topics estimated from the Common Crawl samples compared to the full data. Further, because the subforum samples differ in terms of document count, sample proportion, and other features, we can investigate the relationship between data features and topic representativeness.

The remainder of this paper is organized as follows. We first discuss related work in Section II. In Section III, we provide an overview of the data and discuss our metrics for measuring topic similarity between the sampled and the full data. We discuss our results in Section IV, where we evaluate topic similarity along two metrics: similarity in estimated topic proportions and similarity in word rankings. We further use a multivariate beta regression model to analyze the association between data characteristics and topic representativeness, and we conduct a series of experiments to extend our findings to sample sizes outside our collected data. In Section V, we demonstrate business insights that can be drawn from our estimated topics. Section VI concludes.

II. Related Work

A. Topic Modeling on Online Forums

The value of online forum data has been broadly studied in behavioral research (e.g., [10], [11]). For example, Wu et al. collect data from one of the largest online discussion forums in China to identify the principal users who contribute to a discussion topic [12].

Topic modeling enables research on online forums by identifying underlying topics in forum discussions. Chen et al. use a two tier model to identify popular topics in a large online forum that contains 881,190 posts [13]. The topics identified with a topic model can also serve as data labels because topic models are a form of mixed clustering. Zhou et al. take advantage of the commonly seen Question-and-Answer discussion style in online forums and apply topic modeling to assist the task of suggesting semantically similar questions to a user query [14]. Ramesh et al. use topic modeling to analyze student discussions in three massive open online courses from Coursera [15].

Analysis of vehicle online forums can provide business insights to manufacturers and vendors, such as market structure information. Netzer et al. apply text mining methods to a sedan car forum to estimate sentiment relations between different car models [16]. One finding is that these sentiments are not always explicit and often comprise only a small portion in all forum discussions. That is, car owners generally discuss problems encountered or modifications to their cars without using strong sentimental words. Human tagging is used to evaluate effectiveness of the text mining approach in [16], which is labor intensive and may not be feasible when dealing with massive datasets.

Wu et al. estimate topics from a Honda car owner online forum, which they use to predict how likely a user will participate in a future discussion on a specific topic [17]. They demonstrate that this prediction performs better for regular and active users, and that participation willingness is affected by peer participation in a topic. Shi et al. find that this peer-to-peer relation can rely on other more subtle behaviors, such as browsing [18].

B. Common Crawl Dataset

The Common Crawl data archive [7] is a gigantic public repository of web crawled data, collected and maintained by a non-profit organization "dedicated to providing a copy of the internet to internet researchers, companies and individuals at no cost for the purpose of research and analysis" [19]. Previous research has used the data repository to analyze the graph structure of the web over time [20]. Since a large proportion of the data included in Common Crawl is in the form of text, Common Crawl has also been used in Natural Language Processing (NLP) research, including machine translation ([8], [21]), text classification [22], and taxonomy development [23].

Buck et al., for example, use Common Crawl to build 5-gram counts and language models that improve statistical machine translation [8]. Smith et al. crawl lateral contents of different language pairs from the Common Crawl corpus and use the results to facilitate language translation approaches

[21]. Iyyer et al. use Common Crawl data for evaluating a sentiment classification algorithm based on a deep neural network [22]. Seitner et al. build a tuple database from Common Crawl where each tuple represents a "is-a" relationship between two words, which can be used to analyze more complex taxonomies [23].

C. Comparing Topics

Several methods exist to compare the quality of estimated topics with each other, including perplexity [24], semantic coherence [25], and exclusivity [26]. These methods apply to the comparison of topics estimated on the same data, but using different model parameters (e.g., different number of topics). Our goal is different. We compare LDA topics estimated from two different data sets – Common Crawl and the full data.

To the best of our knowledge, there has been little research on the comparison of topics estimated from different data sources. An exception is [27], which measures the similarity between topics estimated from Twitter and traditional news using the Jensen–Shannon (JS) divergence. This measure compares two topics based on their full word distributions. Our approach (explained in more detail below), in contrast, relies on a set-based comparison between the top keywords from each topic. The top keywords are determined from the word-topic probabilities, but the probabilities themselves are not being compared. We use this measure instead of the JS divergence because our goal is to mimic human evaluation of topic similarity, which would be based on a visual inspection of the top keywords between topics.

III. Method Description

A. Description of Datasets

We choose a very active car owner online forum as our target forum to evaluate Common Crawl's sample quality. This forum is organized into 14 subforums, each representing a different car model made by a specific car vendor. We refer to these by their car type, e.g., "suv-mid" (a mid-sized SUV), "sedan-full" (a full-sized sedan), "convertible-new" (a newer model of a convertible type), etc. For each subforum, we have collected all available data from Common Crawl as well as the full data from the online forum. Since Common Crawl is based on sampled data, it is not an independent dataset from the original forum data (the full data).

The Common Crawl dataset [7] consists of billions of HTML based web pages that are provided in two formats: WARC and WET. The WARC format contains meta data that describes the crawling process, storage hierarchy, HTTP response codes, and HTML tags. The WARC data is more noisy and hence requires filtering and preprocessing before it can be analyzed with LDA. Alternatively, Common Crawl provides extracted raw text data directly for text mining research called the WET format data. However, the WET format data cannot be used for our LDA experiments, since LDA requires detailed separation of texts from different posts and different threads in an online forum. We therefore used the 14 subforum URLs and gathered 280 GB WARC format data files from the publicly available Common Crawl images in AWS. After data preprocessing, we have grouped all posts in a thread together to form one document.

To collect the full data for each subforum, we have developed a customized web crawler based on Jsoup, a java library for working with HTML. Jsoup provides an API for manipulating and extracting data, using the best of DOM, CSS, and jquery-like methods. It can be used to scrape and parse HTML from a URL, file or string. In total, we collected 2.16 TB of raw data. Running the customized crawler on one subforum took, on average, 24 hours. However, a first run of the crawler resulted in the workstation's IP address being blacklisted, which required restarting the data collection with a less aggressive crawling strategy that would decrease the frequency with which the website was accessed. In total, the data collection on the full forum data took over four weeks, while the data collection from Common Crawl was completed in less than one day, which illustrates another advantage of using the sampled data.

We present an overview of the data collected from Common Crawl (CC) and the full data (FD) in Table I, including the number of documents in each subforum and data set, document fraction in the sample compared to the full data, and standard deviation of timestamp gap (in hours) in each subforum. The latter measures variation in the spread of the data over time. Because Common Crawl data is sampled at irregular time intervals, and because subforums differ in their daily user activity, there are differences in the way the data is spread over time between subforums. We capture this variation by first ordering the timestamps of posts in a subforum, then calculating the standard deviation of the intervals between consecutive timestamps of posts.

B. Description of LDA Topic Modeling

Latent Dirichlet Allocation (LDA) [4] is a generative model that estimates latent groups ("topics") from a corpus. Its main assumption is that documents are random mixtures of corpus-wide topics, where each topic is a probability distribution over the entire vocabulary. A key output of LDA is an estimate of each document's topic proportions, which can be used to calculate the proportion of each topic in the entire corpus. We denote these global topic mixtures as $M = \{m_1, \ldots, m_k\}$, where k is the number of topics, and $\sum_1^k m_i = 1$.

For all of the following analysis, we set $k = 10$ and the Dirichlet parameter $\alpha = 0.1$. An exception is the "suv-small" subforum, where we set $k = 5$ because of the small number of documents. We use the original C code provided by [4] available at [28]. On the largest data set in our analysis ("sedan-mid" with 27,649 documents in the full data), the algorithm converges after about eight hours.

Table I
OVERVIEW OF ONLINE FORUMS DATA COLLECTED FROM COMMON CRAWL (CC) AND FULL DATA (FD).

Subforum	Doc. Count CC	Doc. Count FD	Doc. Frac. (CC/FD)	Time-stamp Var.	Size of Raw Data (GB) CC	FD
sedan-mid	3,249	27,649	0.12	38	49.4	267.5
sport-new	1,869	16,553	0.11	33	46.2	216.6
convertible	1,521	28,525	0.05	23	29.7	250.5
suv-compact	1,467	10,187	0.14	30	29.6	128.4
suv-mid	1,397	20,217	0.07	21	17.4	145.8
convertible-new	1,181	8,671	0.14	19	18.2	102.2
sport	901	7,765	0.12	54	19.2	74.4
hatchback	703	3,190	0.22	52	17.2	44.0
sport-full	639	17,110	0.04	34	22.8	493.1
sedan-full	436	3,527	0.12	90	5.3	28.1
coupe-compact	212	15,635	0.01	65	11.3	229.2
electric	193	2,177	0.09	181	7.2	38.2
suv-mid-new	139	9,687	0.01	123	4.2	129.0
suv-small	8	851	0.01	1,014	0.2	13.2

C. Description of Similarity Comparison Between Topics

Our goal is to compare topics estimated from Common Crawl to those estimated on the full data. A common metric for evaluating a topic model's quality is perplexity, which is a measure of a model's predictive likelihood calculated from a held-out set [24]. Perplexity is not a suitable measure in our case because we need to evaluate the similarity of two models estimated on different data rather than comparing their performance on the same data set. We instead use two metrics that capture both the quantitative and qualitative similarity between two topic models. First, we compare topic mixtures estimated from the sampled and full data, using the Kolmogorov–Smirnov (KS) test. Secondly, we compare the top-ranked words in different topics using the Sørensen–Dice coefficient. In this section, we motivate and explain both measures in more detail.

1) Evaluating Topic Proportion Similarity: A key output of LDA is an estimate of the proportion that each latent topic is represented in the corpus, which we denote as M (see Section III-B). An estimate of 0.25 for a topic, for example, means that 25% of the text in a corpus is estimated to fall under this topic. These mixtures are important measures for business analysts because they provide insights into the relative importance of estimated topics.

To give an example, consider the following two topic mixtures estimated for the "suv-compact" forum from Common Crawl (CC) and the full data (FD):

$$M_{\text{CC}} = \{0.16, 0.15, 0.11, 0.11, 0.11, 0.09, 0.09, 0.08, 0.06, 0.05\}$$
$$M_{\text{FD}} = \{0.16, 0.13, 0.11, 0.11, 0.11, 0.09, 0.09, 0.08, 0.07, 0.07\}$$

The topics are sorted from largest to smallest topic for both data sets. This comparison only examines topic mixtures; the semantic alignment of the topics is captured by the other comparison using the Dice coefficient. Based on these mixtures, the two models produce very similar results.

We formally evaluate topic mixture similarity with a two-sample KS test. Let $F_{CC}(x)$ and $F_{FD}(x)$ denote the empirical distribution functions calculated from M_{CC} and M_{FD}, respectively. The two-sample KS test statistics D is then calculated as

$$D = \sup_x |F_{CC}(x) - F_{FD}(x)|. \tag{1}$$

It is a test of the null hypothesis that $F_{CC}(x)$ and $F_{FD}(x)$ come from the same distribution. For the above example, $D = 0.2$ (the largest absolute difference between topic proportions) with $p = 0.99$. Because p is clearly above the standard 0.05 threshold, we cannot reject the null hypothesis that the topic proportions are drawn from the same distribution. While this result does not prove that the two mixtures are the same, there is no statistical evidence that they are different from each other.

The KS test statistic is calculated by matching topics based on their rank. That is, one compares the two largest topic proportions with each other, followed by the second largest proportion, etc. This does not take into account that topics with similar proportions may differ qualitatively, i.e., in terms of their top ranked words that define the topics. We therefore introduce a measure that compares topics qualitatively.

2) Evaluating Topic Meaning Similarity: Topics estimated with LDA are probability distributions over the vocabulary. It is the analyst's job to label the topics, that is, to decide their substantive meaning. This is usually done by sorting the vocabulary by their estimated topic-word probabilities and looking at the top k words, where typical values for k are in the range of 5–20. To provide an example, Table III in Section V, which we will discuss in greater detail below, shows the top ten keywords for the two largest topics estimated from Common Crawl for four selected forums.

For Common Crawl to be a useful sample of the population data, it should produce topics that are substantively similar to those estimated on the full data. We evaluate this criteria with a metric that mimics human evaluation of topic similarity, which would be based on comparing the top keywords of two topics and judging their similarity in terms of the words they include. More precisely, we follow an approach suggested in [29] that uses the Sørensen–Dice coefficient to measure the overlap between two keyword lists. Let X denote the set of k top keywords from a topic estimated from Common Crawl, and let Y denote keywords estimated from the full data. The Sørensen–Dice coefficient (or short, Dice coefficient) is calculated as

$$D(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{2}$$

where $X \cap Y$ is the set of common words from both word lists, and $|X|$ and $|Y|$ are the numbers of words in each list.

To provide an example, consider the following two word lists: $X = \{\text{looks, sport, interior, trim, wheels}\}$ and $Y =$

{wheels, trim, color, price, sport} (these are actual words that appear in the "suv-compact" forum). Each set includes five words, and they share three elements, "sport", "trim" and "wheels". The Dice coefficient for this example is $D(X,Y) = \frac{2\times3}{5+5} = 0.6$. If the two word lists were identical, the Dice coefficient would be 1; if their intersection were empty, the coefficient would be 0. In our calculations below, we will use the same number of the top twenty keywords in each set. The Dice coefficient is then equivalent to the proportion of words that appear in both topics.

The Dice coefficient is a comparison between two topics. When comparing topics from two models estimated on different data sets, we need to assign each topic from one model to a topic from the other model in order to calculate the coefficient. This is a classic matching problem for which well-known solutions exist. We here follow an approach suggested in [29] for this particular case. Because we expect that each topic estimated on the full data has a corresponding topic in the sample, we greedily match topics to each others based on the maximum Dice value. More precisely, for two equally-sized sets of topics, we first match the topic pair with the highest Dice coefficient, then repeat this process with the unassigned topics until all topics are matched. Our measure of model similarity is then the average Dice coefficient over all selected topic pairs.

IV. Results of Topic Similarity

In this section, we analyze the representativeness of topics estimated from Common Crawl compared to those estimated on the full data. We first conduct a comparison between LDA topic proportions between the two data sets, showing that there is no statistical evidence that the topic proportions are not drawn from the same distribution. We then analyze similarity in terms of word ranking, demonstrating that the average Dice values are within a range that would be expected under random sampling in 13 out of the 14 forums. Using a multivariate beta regression model, we show that there is evidence that larger sample proportions and the number of threads in a sample are positively correlated with average topic similarity. Finally, we conduct a series of experiments that generalize our findings to sample sizes not observed in Common Crawl.

A. Topic Proportion Similarity

For each subforum i, we have estimated global topic proportions $M_{\text{CC},i}$ and $M_{\text{FD},i}$. Figure 1 shows scatter plots comparing the two proportions. Most data points are close to the 45-degree line, indicating a high degree of similarity. Their average Pearson correlation is 0.92, with min=0.80 and max=0.97. To formally test the differences between the proportions, we calculate the Kolmogorov–Smirnov (KS) test statistic discussed in Section III for each subforum. These statistics, which are printed in the bottom-right of each panel in Figure 1, range from 0.2 to 0.6, with p-values well above the typical 0.05 threshold. Based on these results, we cannot reject the null hypotheses that the proportions are the same. That is, we do not find statistical evidence that, in terms of their topic proportions, the Common Crawl samples differ from the full data.

B. Word Rank Similarity

We next evaluate the LDA topics estimated from Common Crawl in terms of their substantive similarity with the topics estimated from the full data. Figure 2 shows the average Dice coefficient for each subforum as black dots, ordered from smallest to largest. The average Dice values range from 0.37 ("suv-small") to 0.64 ("suv-compact"), with an average value across subforums of 0.50. In terms of topic similarity, this means that the average matched topic pair between Common Crawl and the full data overlap by, on average, 50% of their top 20 keywords. In 7 out of the 14 subforums, the average Dice value is above 0.5, indicating that the average Common Crawl topic overlaps with more than half of its words with its matched topic from the full data.

We further quantify the similarity comparisons by considering the size of the average Dice value one would expect if the Common Crawl data were a true random sample of the full data. This answers the question to what extent the results estimated on the Common Crawl samples behave the same or differently than under true random sampling. To this end, we conduct the following simulation: for each subforum in the full data, we draw 100 random samples (without replacement) of the same size than the subforum in Common Crawl. For each sample we estimate 10 topics and calculate their average Dice value with the same method we applied to the Common Crawl data. These simulations result in 100 average Dice values for each subforum that correspond to possible results one would obtain under random sampling.

The large black dots in Figure 2 show the average Dice values calculated from Common Craw. The small gray dots show the average Dice value from each of the 100 simulations together with the 95% intervals of estimated values. In 13 out of the 14 subforums, the average Dice value calculated from Common Crawl falls within the 95% interval of Dice values calculated from the random samples. We can conclude that for these 13 subforums, the topic similarity between Common Crawl and the full data is not significantly different than what one would expect to find under true random sampling. The average Dice value is outside the 95% interval in only one case, the "sport-full" forum, which has the fourth smallest sample proportion. This result indicates that samples with a document fraction below 0.05 might not be suitable for topic modeling because of the possibility that the observed sample might be biased.

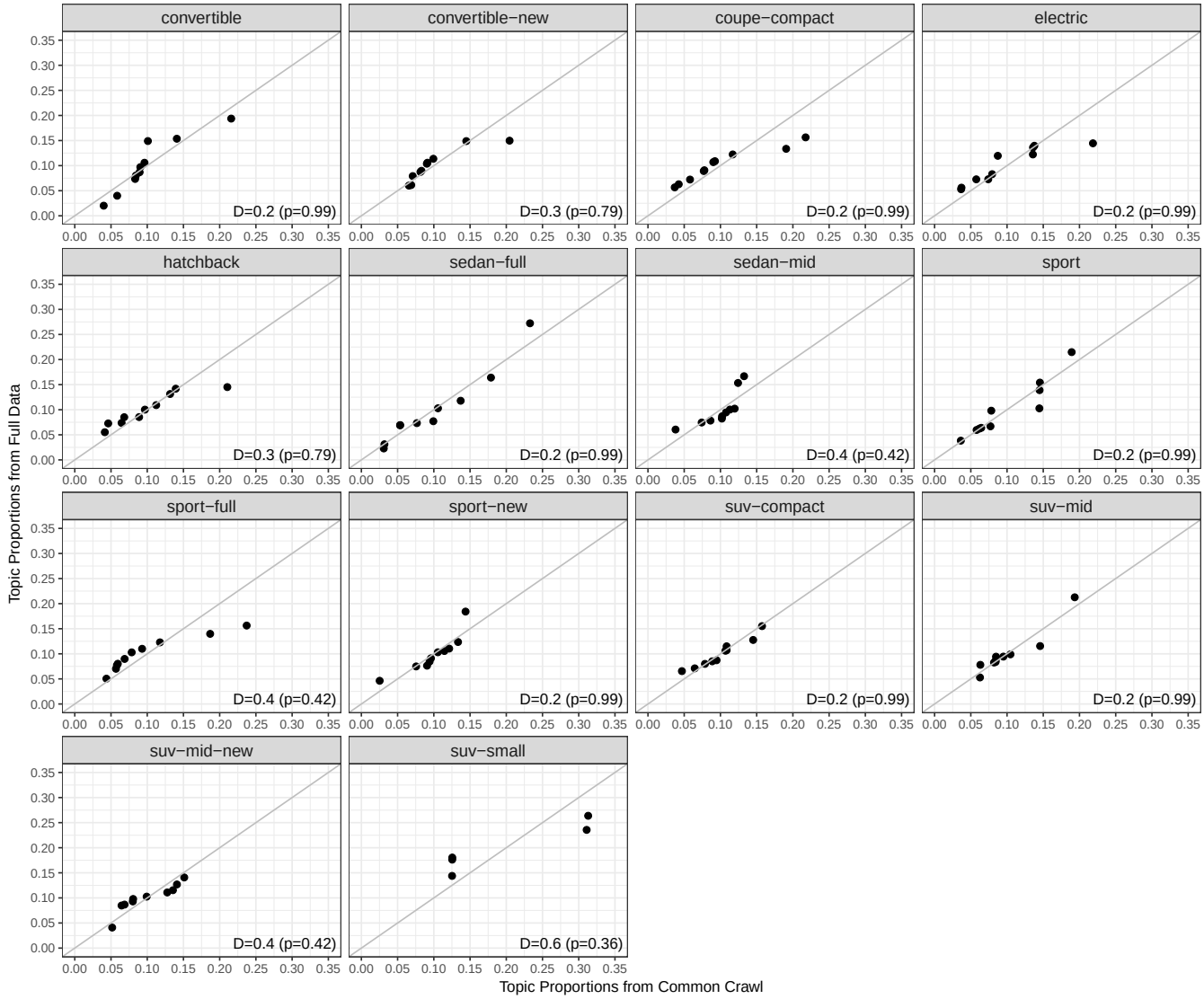

Figure 1. Topic proportions estimated from Common Crawl (x-axis) compared to topic proportions estimated from the full data (y-axis). Gray lines are 45-degree lines. D values in bottom-right are Kolmogorov–Smirnov (KS) test statistics with p-values in parentheses.

C. Multivariate Regression

We have established in the previous section that topic similarity between Common Crawl and the full data does not significantly differ in 13 out of the 14 forums. In this section, we formally test whether differences in data characteristics are systematically linked to topic similarity. We estimate the joint effect of document fraction, document number, and logged time interval variation on average Dice coefficient with a beta regression [30], which accounts for the response being bound in the $(0, 1)$ interval.

Table II shows that document fraction and document number have a positive and significant association with the average Dice coefficient at the 0.1 threshold or below. The estimated coefficients represent additional changes in the log-odds ratio of the response. To facilitate their inter-

Table II
RESULTS FROM BETA REGRESSION

	Coef. (std.dev.)
Document Fraction	1.759^{**} (0.795)
Document Number (1,000s)	0.111^{*} (0.064)
Time interval variation (log)	-0.068 (0.054)
Constant	-0.011 (0.280)
Observations	14
Pseudo R^2	0.643
Log Likelihood	25.563

Note: *p<0.1; **p<0.05; ***p<0.01

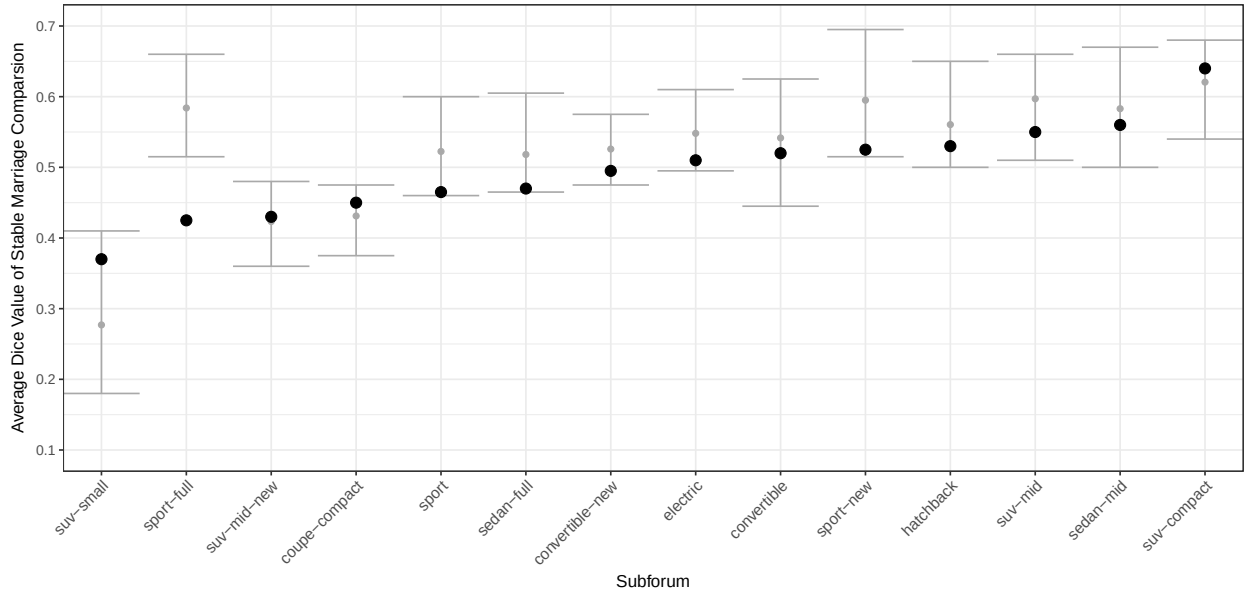

Figure 2. Large black dots are average Dice coefficients calculated from matched topic pairs between Common Crawl and the full data. Small gray dots are average Dice coefficients estimated from 100 random samples drawn from the full data of each subforum and with the same sample size as the Common Crawl subforums. Gray lines indicate the 95% interval of values estimated from the random samples.

pretation, we compute predicted effects when changing a predictor from its empirical minimum to maximum, holding all other predictors at their means. For document fraction, an increase from 0.01 to 0.22 is associated with an average increase in the Dice coefficient by 0.093. For document number, an increase by about 3,200 documents is estimated to increase the response by 0.089.

D. Experiments on Larger Sample Sizes

The largest sample we observe in Common Crawl includes 22% of the full data. We here conduct a series of experiments to evaluate how the quality of topics estimated on sampled data depends on sample sizes outside our observed range. To this end, we draw 100 random samples (without replacement) from the full data of each subforum at six sample sizes: 0.01, 0.1, 0.2, 0.4, 0.6 and 0.8 of the full data. This results in a total of 8,400 samples (14 subforums $\times$ 6 sample sizes $\times$ 100 random samples). For each sample, we estimate 10 LDA topics (and again 5 topics on "suv-small") and calculate the average Dice value between the sampled and the full data. We then calculate the average Dice value and 95% interval range for each set of 100 random samples.

The results are shown in Figure 3. We observe that as a general trend, the average Dice value increases with larger samples. We observe the largest increase when the sample proportion is increased from 0.01 to 0.1 and from 0.1 to 0.2. The line then flattens out at sample proportions above 0.2. The largest average similarity measure we observe in our experiments is 0.71. Considering the full range of values within their 95% intervals, we find a maximum value of 0.83.

The results in Figure 3 provide two important insights for the application of LDA on sampled data. First, we observe that even for samples that include 80% of the full data, the average Dice value does not exceed 0.71 (or 0.83 if we take the full range of values within the 95% interval into account). This indicates that, at least for the data included in our analysis, topics estimated with LDA are sensitive to the types of documents included in the sample. Second, there is a diminishing return in topic similarity for increasing sample sizes. In a majority of cases we observe that topic similarity only increases by a small fraction beyond a 0.2 or 0.4 sample size, indicating that samples at these sizes may be sufficient for the estimation of LDA topics if collecting the full data is too costly or too time intensive.

V. Insights Into Online Forum User Behavior

In this section, we use our estimated topics from Common Crawl to provide insights into customer behavior as expressed in online forum discussions. We focus our discussion on four subforums that we selected because they represent different car classes: "electric", "sedan-mid", "sport", and "suv-compact". Table III shows the top 10 keywords for the two largest topics from each of the car classes.

We observe that for all four classes, the look of the car is the most dominant topic. While the color black is mostly

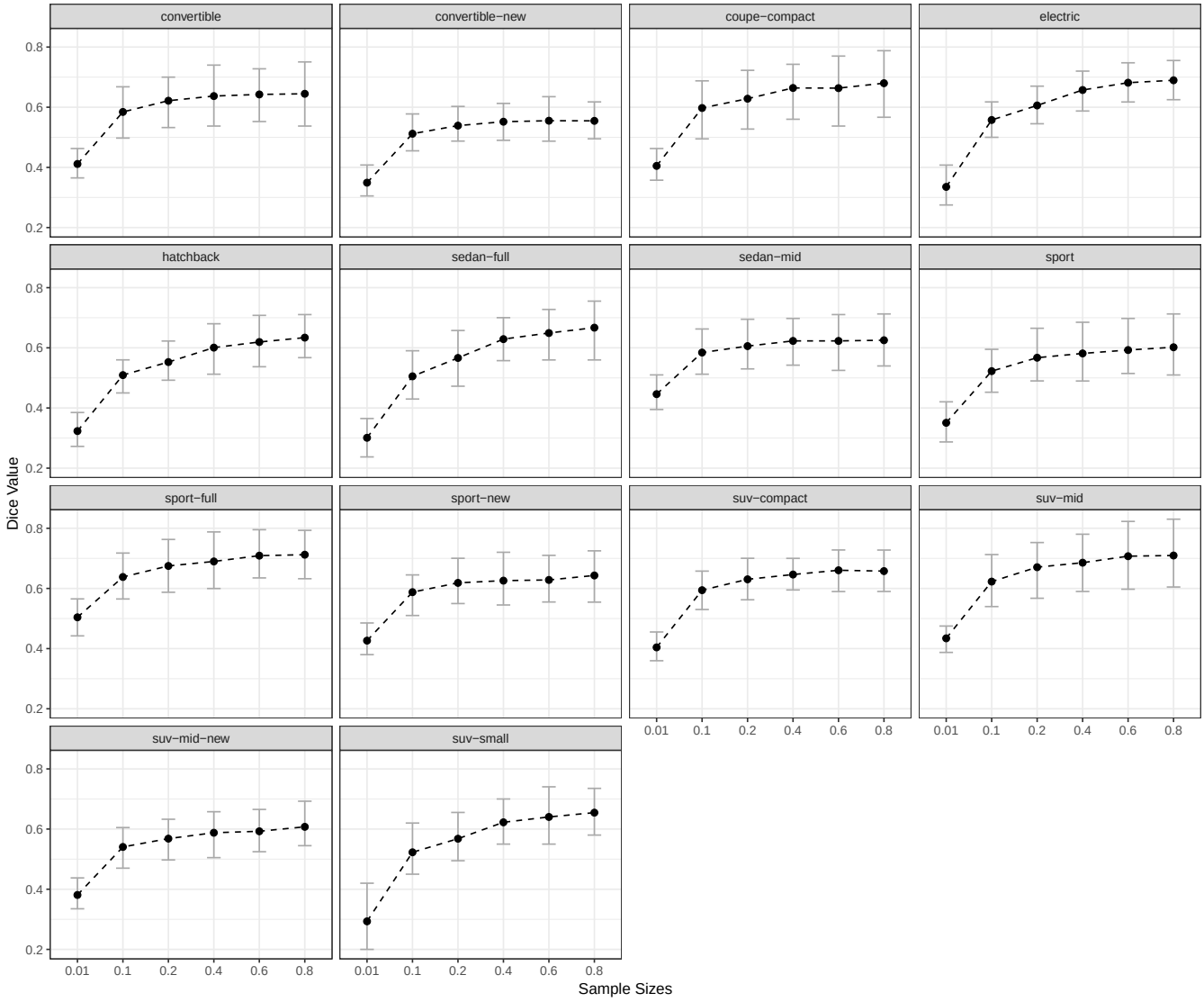

Figure 3. Distribution of the Dice coefficient under random sampling.

mentioned on Sedan-mid, Sport, and SUV-compact classes, the color blue is mostly mentioned on the Electric class.

The second largest topic in each subforum shows distinctive characteristics across the four car classes. In the Electric subforum, the topic represents a focus on battery performance-related keywords covering facets like charging and battery level. In the Sedan-mid class, the second largest topic focuses on phone and iphone associated user experience, suggesting the need for studying compatibility of phone usage inside the car for improved user experience. In the Sport subforum, the second largest topic is around handling and control aspects of the driving experience, covering facets like steering and brakes. Finally, in the SUV-compact subforum, the second largest topic focuses on driving experiences during the winter season, suggesting that the utility of wheels and tires should be of concern for the specific car vendor.

From a business perspective, topic models enable analysts to infer themes from documents in an unsupervised, automated way. By annotating documents with topics, navigation and processing of the text data is improved. Another important application is the combination of topic models and supervised classification approaches, e. g., to sort documents into a fixed set of categories (e. g., a model or defect category). Topic models provide a condensed document representation that is also well suited as input for classification algorithms.

Table III
TOP TEN KEYWORDS FROM LARGEST AND SECOND LARGEST TOPICS FROM FOUR SELECTED SUBFORUMS

Electric		**Sedan-mid**	
First topic	Second topic	First topic	Second topic
car	electric	looks	car
looks	charging	car	vendor
twitter	solar	nice	iphone
electric	battery	sedan-mid	new
blue	level	black	dealer
look	time	look	phone
concept	vehicles	wheels	usb
interior	available	great	need
first	fast	front	system
tesla	use	pics	problem
Sport		**SUV-compact**	
First topic	Second topic	First topic	Second topic
sport	sport	suv-compact	tires
coupe	better	black	suv-compact
sport-new	power	looks	wheels
convertible	performance	sport	winter
nice	weight	interior	snow
love	drive	color	rims
wheels	brakes	pics	suv-mid
black	track	package	need
vendor	steering	trim	price
looks	even	wheels	set

VI. CONCLUSION

In this paper, we have investigated the use of an open source web crawl data repository, the Common Crawl, in LDA topic modeling for online forum data. To evaluate how representative Common Crawl is as a sample for extracting LDA topics, we collected both the full data and the Common Crawl sample for 14 subforums from a car user forum. We compared the LDA topics estimated from Common Crawl samples and on the full data both quantitatively and qualitatively. In both cases, the topics generated from Common Crawl and those drawn from the full data are not statistically different from each other. Through our experiments, we demonstrated that Common Crawl does not perform worse than randomly drawn samples from the full data in terms of topic similarity. We also demonstrated the usefulness of topic models for drawing business insights from an online forum through a discussion of the primary and secondary discussion themes estimated from four representative car classes from the Common Crawl data set.

Our results provide evidence that data collected from Common Crawl is a good candidate for LDA topic modeling on online forums. There are several problems associated with collecting data from online forums directly, including the need to develop a customized web crawler, the possibility of one's IP address becoming blacklisted, the size of the data, and the time required to download the full data. Using Common Crawl as a sample of the full data circumvents many of these problems, and our results show that topics estimated from Common Crawl are not significantly different from the full data in terms of topic proportions, and reasonably similar (and not worse) than under random sampling in terms of word rankings.

Our findings are based on the analysis of 14 subforums that represent different car models made by a specific vendor, but our research strategy provides a template that can be used in other domains to evaluate the representativeness of topic models. Future research will need to investigate the sensitivity of LDA topic modeling results on different online forums and product categories. It is also an open question whether results from extensions of LDA, such as dynamic topic models [31] (which account for topic evolution over time) or hierarchical topic models [32] (which allow for topic hierarchies), would exhibit the same sampling properties. Future work will also explore alternative methods for evaluating the similarity of two inferred topic models, such as by combining the mixture and alignment based metrics used here, or examination of document classification.

Finally, NLP research is increasingly using deep learning systems [33], which are capable of extracting more semantic features from the data. Recurrent neural networks have been proven to capture contextual dependencies. For example, word vector models and deep learning is used to analyze and process textual data. Extensions of our work could investigate if our representativeness estimation can also be applied to these deep learning models.

Code to replicate the data collection and analysis is available at https://www.cs.clemson.edu/dice/.

ACKNOWLEDGMENT

The authors would like to thank Dr. John Holt with LexisNexis Risk Solutions for the suggestion to use the Dice coefficient to compare top-ranked keywords from two topics. This work is funded in part by U.S. Department of Education GAANN award P200A150310 and NSF #1228312.

REFERENCES

[1] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[2] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 2006, pp. 977–984.

[3] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] "Gaia online," http://www.gaiaonline.com/forum/, 2017.

[6] Z. Liu, Y. Zhang, E. Y. Chang, and M. Sun, "Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 26, 2011.

[7] Common Crawl, 2017, last accessed August 17, 2017. [Online]. Available: https://commoncrawl.org/

[8] C. Buck, K. Heafield, and B. Van Ooyen, "N-gram counts and language models from the common crawl." in *LREC*, vol. 2, 2014, p. 4.

[9] A. Stolz and M. Hepp, "Towards crawling the web for structured data: Pitfalls of common crawl for e-commerce," in *Proceedings of the 6th International Workshop on Consuming Linked Data*, 2015.

[10] J. Kropczynski, G. Cai, and J. M. Carroll, "Investigating incidence of common ground and alternative courses of action in an online forum," in *Proceedings of the 15th Annual International Conference on Digital Government Research*, ser. dg.o '14. New York, NY, USA: ACM, 2014, pp. 24–33.

[11] Y. Mou, D. Atkin, H. Fu, C. A. Lin, and T. Lau, "The influence of online forum and SNS use on online political discussion in China: Assessing "spirals of trust"," *Telematics and Informatics*, vol. 30, no. 4, pp. 359–369, 2013.

[12] C. Wu, C. Li, W. Yan, Y. Luo, X. Mao, S. Du, and M. Li, "Identifying opinion leader in the internet forum," *International Journal of Hybrid Information Technology*, vol. 8, no. 11, pp. 423–434, 2015.

[13] F. Chen, J. Du, W. Qian, and A. Zhou, "Topic detection over online forum," in *Web Information Systems and Applications Conference (WISA), 2012 Ninth*. IEEE, 2012, pp. 235–240.

[14] T. C. Zhou, C.-Y. Lin, I. King, M. R. Lyu, Y.-I. Song, and Y. Cao, "Learning to suggest questions in online forums." in *AAAI*, 2011.

[15] A. Ramesh, D. Goldwasser, B. Huang, H. D. Iii, and L. Getoor, "Understanding MOOC discussion forums using seeded lda," *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[16] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko, "Mine your own business: Market-structure surveillance through text mining," *Marketing Science*, vol. 31, no. 3, pp. 521–543, 2012.

[17] H. Wu, J. Bu, C. Chen, C. Wang, G. Qiu, L. Zhang, and J. Shen, "Modeling dynamic multi-topic discussions in online forums." in *AAAI*, 2010.

[18] X. Shi, J. Zhu, R. Cai, and L. Zhang, "User grouping behavior in online forums," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 777–786.

[19] Common Crawl website, "Frequently Asked Questions," 2017, last accessed August 17, 2017. [Online]. Available: https://commoncrawl.org/big-picture/frequently-asked-questions/

[20] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer, "Graph structure in the web—revisited: a trick of the heavy tail," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 427–432.

[21] J. R. Smith, H. Saint-Amand, M. Plamada, P. Koehn, C. Callison-Burch, and A. Lopez, "Dirt cheap web-scale parallel text from the common crawl." in *ACL (1)*, 2013, pp. 1374–1383.

[22] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the Association for Computational Linguistics*, 2015.

[23] J. Seitner, C. Bizer, K. Eckert, S. Faralli, R. Meusel, H. Paulheim, and S. Ponzetto, "A large database of hypernymy relations extracted from the web," in *Proceedings of the 10th edition of the Language Resources and Evaluation Conference, Portoroz, Slovenia*, 2016.

[24] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1105–1112.

[25] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[26] J. Bischof and E. M. Airoldi, "Summarizing topical content with word frequency and exclusivity," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012, pp. 201–208.

[27] X. Zhao and J. Jiang, "An empirical comparison of topics in twitter and traditional media," *Singapore Management University School of Information Systems Technical paper series. Retrieved November*, vol. 10, p. 2011, 2011.

[28] David Blei, 2017, last accessed August 19, 2017. [Online]. Available: https://github.com/blei-lab/lda-c

[29] C. Gropp, A. Herzog, I. Safro, P. W. Wilson, and A. W. Apon, "Scalable dynamic topic modeling with clustered latent Dirichlet allocation (CLDA)," *arXiv preprint arXiv:1610.07703*, 2016.

[30] F. Cribari-Neto and A. Zeileis, "Beta regression in R," *Journal of Statistical Software*, vol. 34, no. 2, pp. 1–24, 2010.

[31] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 2006, pp. 113–120.

[32] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *Journal of the ACM (JACM)*, vol. 57, no. 2, p. 7, 2010.

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.