Corporate Social Responsibility (CSR): A Survey of Topics and Trends Using Twitter Data and Topic Modeling

Bongsug (Kevin) Chae 1,* and Eunhye (Olivia) Park 2

1 Department of Management, Kansas State University, Manhattan, KS 66506, USA
2 Department of Hospitality Management, Kansas State University, Manhattan, KS 66506, USA; eunhyepark@ksu.edu
* Correspondence: bchae@ksu.edu; Tel.: +1-785-532-3185

Received: 27 May 2018; Accepted: 26 June 2018; Published: 28 June 2018

Abstract: Corporate social responsibility (CSR) is an essential business practice in industry and a popular topic in academic research. Several studies have attempted to understand topics or categories in CSR contexts and some have used qualitative techniques to analyze data from traditional communication channels such as corporate reports, newspapers, and websites. This study adopts computational content analysis for understanding themes or topics from CSR-related conversations in the Twitter-sphere, the largest microblogging social media platform. Specifically, a probabilistic topic modeling-based computational text analysis framework is introduced to answer three questions: (1) What CSR-related topics are being communicated in the Twitter-sphere and what are the prevalent topics or themes in CSR conversation? (topic prevalence); (2) How are those topics interrelated? (topic correlation); (3) How have those topics changed over time? (topic evolution). The topic modeling results are discussed, and the direction for future research is presented.

Keywords: corporate social responsibility; topic modeling; Twitter; sustainability

1. Introduction

Corporate Social Responsibility (CSR) has a long history in academic literature, beginning in the 1950s [1]. The concept experienced significant growth in importance and interest through the 1980s and 1990s [2]. Today, CSR is considered a major topic in academic literature and an essential business practice across industries, along with associated terms such as sustainability and sustainable development. Several authors have offered comprehensive reviews of CSR research and business practices [2–5]. For example, Aguinis and Glavas [3] show that CSR literature has focused on three broad areas, including predictors of CSR (why), mediators and moderators of CSR-outcomes (how), and outcomes of CSR. Wang et al. [4] identified four major trends in CSR research: more process-oriented studies, growing attention to organizational performance, focusing on specific dimensions of CSR, and increasing non-U.S.-based studies.

Others also have presented definitional/theoretical frameworks for understanding CSR as a business practice [6–8]. In this line, Carroll [6] discussed a pyramid of CSR, including four dimensions: economic (be profitable), legal (obey laws and regulations), ethical (be fair and just), and philanthropic (be a corporate citizen). Economic and legal dimensions are considered requirements for a firm’s survival, while the firm is expected to be ethical and desired to be philanthropic. According to business executives [9], there are different weights on each of the dimensions—economic, legal, ethical, and philanthropic, in order of importance. In a similar vein, Garriga and Melé [8] classified extant CSR approaches into four dimensions: instrumental, political, integrative, and ethical.
While a literature review clearly indicates that CSR is increasingly studied in academic research, and its adoption continues in industries and countries [4], there are many needs and opportunities for future CSR research. One potential area is the application of content (or discursive) analysis in CSR research [10,11]. Content analysis aims to extract valuable insights (e.g., themes, categories, sentiments) from text corpora, such as academic articles, website contents, and industry reports. More recently, authors in diverse academic fields have paid attention to information retrieval from diverse CSR-related text sources, including published academic articles, websites, corporate reports (e.g., 10k, sustainability reports), and newspapers.

In this vein, Dahlsrud [7] analyzed 37 popular definitions of CSR in the literature and, using content analysis, identified five dimensions from those definitions: environmental (natural environment), social (the relationship between business and society), economic (economic or financial aspects), stakeholder, and voluntariness. In addition to academic articles, CSR reports have been a popular source of research data and authors have applied content analysis. Tate et al. [12] analyzed CSR reports of 100 companies and reported on major themes in their writing. Popular themes are “community focus,” “consumer orientation,” “risk management,” “energy,” and “health.” Lee and Carroll [13] examined the changes to CSR in the public sphere over 25 years (1980–2004), by applying content analysis to opinion pieces (e.g., editorials, letters-to-the-editor, op-eds, and guest columns) of nine U.S. newspapers. Their analysis of 460 opinion pieces shows that the four dimensions of CSR [9] have changed their prominence in newspapers: legal and philanthropic issues were popular in the late 1990s, while economic and ethical issues emerged as important in the early 2000s.

These existing studies using content analysis and text corpora have tended to examine relatively small samples of text data from traditional information channels (e.g., journal articles, CSR reports, newspapers). While social media is increasingly important for CSR communication and information sharing [14,15], the use of social media data is rare in the literature. Also, for analysis, the previous studies have used human coding and/or rather simple, software-based, coding methods [11] in their content analysis.

This study adopts computational content analysis [16,17] for understanding themes or topics from CSR-related conversations in the Twitter-sphere, the largest microblogging social media platform. Computational content analysis, a big data analytics method, is suitable to process big text data and discover hidden patterns in the data. Specifically, a probabilistic topic modeling-based computational text analysis framework is introduced to answer three questions:

1. What are CSR-related topics being communicated in the Twitter-sphere? What are the prevalent topics or themes in CSR conversation? (topic prevalence)
2. How are those topics interrelated? (topic correlation)
3. How have those topics changed over time? (topic evolution)

First, the extant literature on CSR dimensions or themes indicates that no single dimension or activity is enough to define CSR. Instead, CSR involves multiple dimensions, and several social activities are exercised under the umbrellas of CSR. Thus, Question #1 employs a machine learning-based content analysis method, called “Structural Topic Modeling” (STM) [17], which is suitable for information retrieval from big data. For this study, 1,286,668 tweets were collected, over a three-year period, between October 2013 and November 2016. Second, while the dimensions (or topics) of CSR are analytically distinct, they are not independent of each other, but rather interrelated. For example, sustainable product development is not separable from energy or climate issues. In turn, Question #2 examines how different CSR topics are correlated. Third, CSR is an evolving construct [1] in institutional contexts. Its dimensions (or topics) have evolved over time. Some dimensions are more prominent than others during a given period. Question #3 addresses how different dimensions have changed between October 2013 and November 2016.

This study makes a three-fold contribution to the community of CSR research and practice. First, the study identifies social media data as the source of understanding CSR phenomenon. As noted
earlier, the extant research has recognized the value of “texts” as data, but the focus has been on traditional communication channels. Social media data is large in volume and is readily accessible by researchers and practitioners using Application Programming Interface (API). Second, this study illustrates how computational content analysis and, in particular, probabilistic topic modeling can be used as an effective research method in managing big text data and automatically discovering latent structure. Finally, the study proposes a framework showing how social media data and computational content analysis can be operationalized in studying CSR. The proposed framework can be useful for future research and industrial applications.

2. Literature Review

2.1. Computational Content Analysis

Content analysis has emerged as a useful tool for conducting CSR research. Content (or discursive) analysis broadly refers to a systematic approach to extract meaningful information from text documents, including academic papers, corporate reports, company website pages, newspapers, and social media posts. Content analysis is growing in popularity among CSR studies [11]. A popular content analysis uses human reading and coding of text corpora. Several studies have adopted this approach to analyze newspapers [13,18], academic articles [10,19], company web pages [20,21], and company reports [22]. Another popular method of content analysis relies on some sort of software coding for discovering meaningful structure from text data [11]. Tate et al. [12] used a software (“Crawdad”) in identifying themes related to CSR and supply chain management (SCM) from companies’ CSR reports. These approaches are also used together in CSR research [23].

While having been useful for CSR research, the popular content analysis approaches alone are less effective for the analysis of big data, which may include millions (or billions) of rows of text. Today, corporate reports and sustainability disclosures are increasingly available in digital formats (e.g., GRI Initiative, CorporateResister, CSwire, CSRHUB). Also, large volumes of CSR-related communication and information flow through social media platforms (e.g., Twitter, Facebook, blogs) [24–26]. As CSR-related data increases in volume, computational text analysis emerges as an effective tool for quickly processing, visualizing, and analyzing such large CSR datasets.

Computational text analysis uses natural language-processing techniques (NLP) and advanced machine learning (ML) algorithms, which are often borrowed from computer science and statistics. Text data, in general, is known to be unstructured and messy, compared to traditional data from corporate relational databases. NLP offers an effective tool for cleaning and transforming text data for content analysis, including simple descriptive text analysis (e.g., word frequency, word cloud, word association analysis), and advanced content analysis.

Two types of ML-based approaches are considered for advanced content analysis: supervised and unsupervised. To use supervised ML-based text analysis, researchers need the text corpora containing known categories or labels, and to build predictive models using such algorithms as Support Vector Machine (SVM) and Naïve Bayes. Text classification (e.g., if a stakeholder’s Facebook post about a company’s CSR is positive or negative) is a popular application of supervised ML. The key challenge with the supervised approach is that known categories (or a training dataset) may not be readily available in the research data and preparing a training dataset to build the predictive models, especially for big data analysis, takes a great deal of time.

On the other hand, unsupervised ML approaches do not need the data with known categories (a training dataset). The unsupervised text analysis takes raw texts as the input, preprocesses and transforms them through NLP techniques (e.g., removing stopwords, converting words to numbers) and, finally, attempts to discover categories or topics from the text data using advanced statistical algorithms. This unsupervised approach can be scaled to large datasets and has appeared attractive to social science researchers [27,28]. For example, traditional clustering algorithms (e.g., k-means) reveal similarities between documents or texts [29], and recent algorithms, such as Latent Dirichlet
Allocation (LDA) [16] and Correlated Topic Model (CTM) [30], uncover latent topics from large amounts of text data [31].

2.2. Topic Modeling for Business Research

Topic modeling is an unsupervised machine learning-based content analysis technique focusing on automatically discovering hidden latent structure from large text corpora. In topic modeling, a document is considered a collection of words containing multiple topics in different proportions. For example, a social media post may be largely about natural environments, while it is also about health and supply chain (i.e., 70% natural environments, 20% health, and 10% supply chain). A topic (or theme) means a list of semantically coherent words in different weights. Thus, the “health” topic may be represented by such words as health (40%), tobacco (10%), safety (20%), risks (20%), and addiction (10%). In topic modeling, the latent structure refers to two types of information from the data: document-topic distribution and topic-term distribution. The document-topic distribution informs as to how each document (or social media post) is composed, in terms of topics. The topic-term distribution provides different lists of semantically coherent words, where each list of terms represents a topic or theme.

Latent Dirichlet Allocation (LDA) [16] is the most popular topic model discussed in the literature. The way LDA operates is by asking the following question: “What is the latent structure that likely generated the observed collection?” (p. 79). The documents and the words ($w$) in those documents are observed, but the topic distributions for each document ($\theta$), the topic assignment for each word in each topic ($z$), and the term distributions over topics ($\beta$) are considered latent (or hidden) structures (Figure 1). This leads to a generative process whose goal is to discover the topic structure ($\theta$, $z$, and $\beta$), which can explain the observed word ($w$). Figure 1 graphically represents this generative process.

![Figure 1. A graphical representation for LDA (Adapted from Blei 2012).](image)

LDA is simple and modular and, thus, has been adapted to develop more sophisticated topic models [32] for specific purposes in research and industry practices, including the correlated-topic model and author-topic model. Many of these LDA-based topic models are developed in the computational science fields (or the machine learning community), where the focus is on discovering the overall topics from big text data. On the other hand, social scientists (and business researchers) often have additional information (or metadata) about documents. For example, a corporate sustainability report provides (or can be linked into) meta-data, which includes company name, company financial data (e.g., revenues, credit rating), and other company- and industry-related data (e.g., governance structure). These covariates are important in business research when exploring CSR-related conversations.

Structural Topic Modeling (STM) [17,33] is a relatively new probabilistic topic model, incorporating covariates or additional document-level information in the process of inferring topics. This makes STM very suitable for computational social science (or business) research using text data. Specifically, STM adds two additional components to the LDA model: topic prevalence and topic
content (Figure 2). Topic prevalence allows such numerical covariates \( (X) \) as report year, company size, and revenues, to influence the topic proportion \( (\theta) \). This is done by applying a logistic normal linear prior \( (\theta \sim \text{LogisticNormal} (X)) \), instead of a Dirichlet prior \( (\theta \sim \text{Dirichlet} (a)) \). Thus, for example, researchers can examine whether topic prevalence (e.g., environmental dimension) has changed over the years \( (X) \). In this case, the covariate variable “year” \( (X) \) affects per-document topic proportions \( (\theta) \). Using a multinomial logit function, STM also allows researchers to test whether categorical covariates \( (Y) \) affect per-topic term distributions \( (\beta) \).

STM is fully implemented as an R package, and the package offers features not only for topic discovery, but also for text preprocessing, model search and validation, and topic visualization [17]. Figure 3 illustrates the process of topic modeling using STM, which will be used in our study.

**Figure 2.** A graphical representation of STM (Adapted from Roberts et al. 2014).

**Figure 3.** The topic modeling process using STM.

Topic discovery focuses on identifying topics from CSR-related conversations in the Twitter-sphere and assigning a suitable label for each topic. This has been the primary goal of topic modeling in the literature. Topic visualization often involves representing topics using word clouds and other data visualization techniques. Topic correlation is primarily interested in finding interrelated topics, but researchers can also identify CSR dimensions while viewing interrelated topics such as a CSR
dimension or a major category (or a level higher than individual topics). Topic evolution considers topics as dynamic entities changing in popularity over time.

3. Materials and Methods

3.1. Data (Text Corpus) and Text Preprocessing

Twitter was chosen as the source of data for this study because the social media platform is popular among businesses and industry professionals for information sharing, marketing, and engagement e.g., [34]. Twitter provides Application Programming Interface (API), which allows researchers to collect all tweets with specific keywords or hashtags in real time. Python was used for data collection and text preprocessing. The data collection using hashtag #csr through Twitter API, between October 2013 and November 2016, resulted in harvesting over 1.2 million Twitter posts. First, we removed non-English tweets and retweets, and this led to 570,499 English tweets. Text preprocessing involves removing urls, usernames, unnecessary characters, and numbers from the dataset. Short texts such as tweets introduce the challenge of data sparsity because probabilistic topic modeling uses words’ co-occurrence [35]. To handle data sparsity, we considered four additional steps of text processing. First, we “grouped” the preprocessed tweets by username and per week. This resulted in 214,628 expanded tweets. Second, a language identification Python package (“langid”) was applied to remove further non-English tweets. Third, any tweet with fewer than four words was removed. Finally, tweets with fewer than 40 characters were removed because of potential data sparsity issues associated with short texts. The final dataset contained 178,908 expanded tweets.

3.2. Model Section and Topic Modeling Using STM

The STM R package was used for model selection and topic modeling. Topic modeling is an unsupervised content analytics algorithm, and there is only one input—the number of topics (k) being searched—required from researchers. The decision for the optimal k should be made, because it is possible that many topic models can be constructed with different k values in the same corpus. To select the optimal k value, there are different techniques—cross-validation, residuals, semantic coherence—, which is the model selection process [17,32,36]. These methods vary in terms of interpretability, scalability, and other measurements. This study used the mixed methods to determine the optimal number of topics (k). First, the model fit was estimated by comparing the residuals of the models with the different k (from 2 to 80). Figure 4 shows the result of an analysis of residuals [37] available in the STM R package. The result demonstrates the best model fit when k is between 30 and 50.

![Figure 4. The results of residual-based model selection.](image-url)
Second, for those models with low residuals, researchers compared the quality of individual topics in two dimensions: topic coherence and topic exclusivity. A topic is considered cohesive when top words representing the topic tend to co-appear in a corpus, meaning the topic (e.g., philanthropy) is represented by semantically coherent words (e.g., giving, causal marketing, nonprofit, love). A topic can be said to be "exclusive" when top words representing the topic do not appear as top words for other topics [17]. Thus, topic exclusivity complements topic coherence in terms of evaluating overall topic quality. Detailed computational algorithms for these two criteria are available in STM documentation (The function "topicquality" is available in STM R package to visualize topic coherence and exclusivity for each topic https://rdrr.io/github/bstewart/stm/man/topicQuality.html). Based on these two criteria, researchers agreed that the topic model performed best when the number of topics was 31.

We then examined the popular terms of each topic. This showed that four topics (9, 17, 21, and 31) were not related to CSR (Corporate Social Responsibility). Instead, Topics 9, 21, 31 are about customer service representative (CSR) and related jobs and hiring. Topic 17 is about a mobile app called CSR2 (car racing game). Thus, we excluded these four topics from further discussion.

4. Results

4.1. Topic Discovery and Visualization

For topic discovery, we relied on two outputs of STM topic modeling (per-topic word distribution and per-document topic distribution). The per-topic word distribution offers top words for each topic, so this has become the primary source of topic discovery in developing a suitable label for each topic. We also used the per-document topic distribution to find tweets most relevant to each topic and this information was used as complementary in labeling topics.

After reviewing the per-topic word distribution (the list of top words per topic), we assigned a label for each topic. For example, top words for Topic 1 are company, value, strategy, key, profit, society, create, and stakeholder and, thus, Topic 1 was labeled "company strategy." Topic 6 included top words such as sustainable, supplychain, supply, industry, chain, and transparency, leading to "supply chain" for the topic label. Based on top words such as philanthropy, giving, nonprofit, back, corporation, love, action, and causemarketing, we labeled Topic 12 "philanthropy." Topic 15 is labeled "community & charity," based on such top words as community, charity, initiative, project, and local. Also, we labeled Topic 20 "Human rights," based on top words like right, humanrights, human, pay, and living. Topic 30 was labeled "entrepreneurship," based on top words such as enterprise, environmental, entrepreneurship, and citizenship.

For illustration purposes, a plot was created to show shared words and distinctive words between Topic 1 and Topic 15. Figure 5 indicates that (1) Topic 1 is about business and strategy, while Topic 15 is about charity for community and education, and (2) Topic 5 is about health-related issues, and Topic 6 is about sustainable supply chain. Topic 1’s top words are company and strategy, which are quite distinct from Topic 15’s top words such as community, charity, and partner. Likewise, there are distinctive words, not shared words, between Topic 5 and 6, indicating these topics are not related (more discussion in Section 4.2 topic correlation).

Also, since this study analyzed Twitter posts, there were topics related to CSR events, announcements, jobs, reports, and conferences. For example, Topic 4 (help, need, want, know, ngo, etc.) is largely about NGO volunteering and help wanted. Topic 10 has top words such as great, work, team, and woman and is labeled CSR team. Topic 26 is related to CSR conferences (learn, look, conference, video, etc.). Topic 27 is represented by top words such as report, launch, global, annual, and release (CSR reports).

Labeling some topics was relatively challenging because top words alone did not provide a clear picture about the topic. For example, top words for Topic 22 were business, socent, innovation, sharevalue, and socialbusiness. According to the per-document topic distribution, some relevant
tweets are “The priorities for business are challenged now. #socialbusiness #socent #csr … profit versus purpose? Why the two do not need to compete with each other #socent #socialbusiness #csr”. We labeled the topic as “social business.” Table 1 shows sample tweets for each topic with their document-topic distributions (weights).
<table>
<thead>
<tr>
<th>Label</th>
<th>Top Words</th>
<th>Original Tweet</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 Company strategy</td>
<td>company, value, strategy, key, profit</td>
<td>[#Webinar] The future of reporting: How to build effective #stakeholder dialogue to boost your #CSR strategy</td>
<td>0.64</td>
</tr>
<tr>
<td>T2 India CSR</td>
<td>india, effort, company, list, congrats</td>
<td>Status of #CSR spending by Indian Company or foreign companies spend.@nsitharaman @runjaitley @FinancialXpress</td>
<td>0.67</td>
</tr>
<tr>
<td>T3 Green initiatives</td>
<td>sustainability, green, training, gri, future</td>
<td>Sustainability Purchasing course—SustainTrain Online #CSR #Sustainability Courses #elearning #csr</td>
<td>0.95</td>
</tr>
<tr>
<td>T4 Help needed</td>
<td>help, need, want, know, like</td>
<td>want to help #merchants #clients in a #CSR role in #LibertyLake #WA or #CoralSprings #FL #imhiring</td>
<td>0.98</td>
</tr>
<tr>
<td>T5 Health, tabacco</td>
<td>health, tobacco, advertising, reduced, addiction</td>
<td>Have restrictions on tobacco #advertising reduced tobacco #addiction? #Health #CSR #DevelopingCountries</td>
<td>0.96</td>
</tr>
<tr>
<td>T6 Supply chain</td>
<td>sustainable, supplychain, supply, industry, chain</td>
<td>Global retailers shrimp supply chains linked to Thai slave factory #supplychain #CSR</td>
<td>0.74</td>
</tr>
<tr>
<td>T7 Employee engagement</td>
<td>employee, program, volunteer, engagement, startup</td>
<td>3 Awesome Reasons You Should Consider a #Corporate #Wellness Program #CSR #startup #bizstrategy</td>
<td>0.87</td>
</tr>
<tr>
<td>T8 CSR award meeting forum</td>
<td>award, sector, last, group, meeting</td>
<td>The Peer Awards for Excellence in #CSR, #HR; #CustomerEngagement deadline for entry: 01/05</td>
<td>0.87</td>
</tr>
<tr>
<td>T10 CSR teams</td>
<td>great, work, team, woman, working</td>
<td>@AAUW; @projectsgirls work to close the gender gap in #STEM fields by inspiring girls in #ComputerScience #CSR</td>
<td>0.57</td>
</tr>
<tr>
<td>T11 Vegan, meatless</td>
<td>check, read, give, best, come</td>
<td>MeatFreeMonday Try this #MeatlessMonday with #Vegan Roasted Cauliflower and Spiced Chickpea Salad #GlutenFree</td>
<td>0.90</td>
</tr>
<tr>
<td>T12 Philanthropy</td>
<td>philanthropy, giving, nonprofit, back, corporation</td>
<td>Is your #corporation giving back? We’d love to cover your #CSR and #philanthropy. Email us info at <a href="mailto:missa@askmissa.com">missa@askmissa.com</a> so we can cover!</td>
<td>0.98</td>
</tr>
<tr>
<td>T13 Leadership trust culture</td>
<td>leadership, trust, tchat, hrblogs, culture</td>
<td>Is Fair Treatment of Employees Essential for Engagement? MY LATEST! #hrblogs #hr #hr247 #CSR Story</td>
<td>0.99</td>
</tr>
<tr>
<td>T14 Giving crowdfunding</td>
<td>one, socialgood, gift, child, crowdfunding</td>
<td>Donate to (*person name) JustGiving page #cancer #treatment #prevention #justgiving #supportcancer #CSR #risk</td>
<td>0.92</td>
</tr>
<tr>
<td>T15 Community charity</td>
<td>community, charity, initiative, project, education</td>
<td>#CSR Raise awareness of tourism and boost tourism entrepreneurship by sponsoring members of local communities to enroll in our course.</td>
<td>0.97</td>
</tr>
<tr>
<td>T16 Corporate governance</td>
<td>esg, sustainability, corpgov, trend, sri</td>
<td>Invigorating orgs to move forward to specific milestones on their journey to #success. #OrgDev #CSR</td>
<td>0.96</td>
</tr>
<tr>
<td>T18 Climate energy water waste</td>
<td>change, climate, energy, water, food</td>
<td>Tree planting using drones this could be an ideal way to carry out Corporate Social Responsibility #CSR</td>
<td>0.94</td>
</tr>
<tr>
<td>T19 Public benefit</td>
<td>benefit, public, product, london, cost</td>
<td>Featured #CSR Article: Tackling waste in the global food supply chain</td>
<td>0.83</td>
</tr>
<tr>
<td>T20 Human rights</td>
<td>right, humanrights, human, pay, living</td>
<td>#CEO pay surges in #Minnesota, widening gap with workers @StribBiz #Compensation #ExecutivePay #CSR</td>
<td>0.73</td>
</tr>
<tr>
<td>T22 Social business</td>
<td>business, socent, innovation, doe, biz</td>
<td>Why it’s vital to ask yourself WHY BOTHER doing what you do, in business. #SocEnt #CSR #HR #business vision statement should INSPIRE and MOVE you.</td>
<td>0.91</td>
</tr>
<tr>
<td>T23 Top latest story</td>
<td>top, latest, story, see, rse</td>
<td>The latest Achieve More Together Daily! Thanks to @nickfassler @bethanyfhogan @DianeRemin #CSR</td>
<td>0.87</td>
</tr>
<tr>
<td>T24 CSR cases</td>
<td>susty, make, making, case, difference</td>
<td>Read @Giles_Cross in #SanlamSpeak talking #CSR and #responsibility #SanlamPeople making a difference</td>
<td>0.82</td>
</tr>
<tr>
<td>T25 Business ethics</td>
<td>responsible, global, leader, ethic, consumer</td>
<td>What Does #Anticorruption have to do with #CSR?</td>
<td>0.76</td>
</tr>
</tbody>
</table>
Figure 6 shows the topic labels, as well as the expected topic proportions, which represent the popularity of each topic in the Twitter-sphere. Given that 31 topics were identified in this study, the average topic proportion per topic is about 0.032 (100%/31 = 3.2% − average topic proportion). There are noticeable differences in the 31 topics in terms of topic prevalence (or popularity).

Between 2014 and 2016, popular topics were related to company strategy in CSR contexts (Topic 1), community charity (Topic 15), climate and energy-related issues (Topic 25), supply chain (Topic 6), and corporate/environmental/social governance (Topic 16). This indicates these topics were frequently discussed by Twitter users. Some less prevalent topics included Topic 5 (health, tobacco, advertising, developing countries), Topic 14 (gift, giving, crowdfunding, giveback), Topic 13 (leadership, trust, culture, management), Topic 19 (benefit, public, product, cost), and Topic 20 (human rights, pay, living). Word clouds were created to show top words per each topic (Appendix A).
4.2. Topic Correlation

Network visualization was constructed to understand how different topics are related to each other (Figure 7). Nodes and node sizes are topics and their expected topic proportions respectively. For example, the node sizes of Topic 1 (company strategy) and Topic 15 (community charity) are bigger than other nodes. The width of edges is based on correlation analysis, which was performed on the document-topic distribution ($\theta$): if two topics (e.g., company strategy, business ethics) tend to co-appear in tweets, this increases the correlation between the two topics. The coefficient score 0.05 was used as the threshold, so the edges whose correlation coefficients are below 0.05 were removed from the network.

The topic correlation network shows that 20 topics appear to be associated with one or more topics, and seven topics are without statistically significant correlation with other topics. Topical communities are formed from those 20 topics. The largest community, represented by 13 of those 20 topics, consists of some of the topics popular in academic literature and industry press, including Topic 1 (company strategy), Topic 6 (supply chain), Topic 16 (corporate governance), Topic 22 (social business), Topic 25 (business ethics), and Topic 30 (entrepreneurship), among others. The central node of this community is Topic 25 (business ethics), which links popular topics such as Topic 1 (company strategy) and Topic 22 (social business), Topic 5 (supply chain), and Topic 20 (human rights). In contrast, the other community is formed by topics that do not popularly appear in typical academic studies, for events such as community charity (Topic 15), vegan meatless (Topic 11), and help needed (Topic 4).

![Figure 7. Topic Correlation and Community Detection in Topic Network.](image)

4.3. Topic Evolution

To study the interest in CSR topics over time, the effect of time as a covariate on the topics was plotted. Figure 8 shows such plots indicating the changes of topic popularity between 2013 and 2016. The middle line in each plot shows how topic proportions vary over time as a continuous variable, and the other two dotted lines show the confidence interval at 0.95.
The plots show topic proportions have fluctuated over the years. However, there are also some evolutionary patterns (e.g., growth, decline) for several topics. For example, some topics have grown in popularity, including Topic 7 (employee engagement), Topic 15 (community charity), and Topic 23 (CSR latest story). Also, topics such as Topic 5 (health, tobacco) and Topic 12 (philanthropy) have declined in topic proportions.

Based on the results of topic correlation and topic evolution, we further performed trend analysis for some exemplar topic clusters, and the results were illustrated in Figure 9. The following formulae were used for the analysis:

Economic cluster = Topic 1 (Strategy) + Topic 22 (Social business) + Topic 30 (Entrepreneurship)
Environmental cluster = Topic 3 (Green initiative) + Topic 18 (Climate energy water waste)

Philanthropic cluster = Topic 4 (Help needed) + Topic 14 (Crowdfunding) + Topic 15 (Community charity)

Ethical cluster #1 = Topic 1 (Strategy) + Topic 22 (Social business) + Topic 25 (Business ethics)

Ethical cluster #2 = Topic 6 (Supply chain) + Topic 20 (Human rights) + Topic 25 (Business ethics)

Emerging economies cluster = Topic 2 (India) + Topic 28 (Saudi Arabia)

Figure 9. Evolution of Topic Clusters.

The topics of three clusters (economic, philanthropic, emerging economies) appear to be increasing in trend. The topics of the philanthropic cluster are community-oriented as shown by the top words of those topics, such as help needed, ngo, crowdfunding, give back, gifts, community, charity, donation, and local. There is a strong positive trend of this community focus cluster. The other three clusters are either steady (ethical cluster #1) or declining (environmental cluster, ethical cluster #2). Since Topic 25 (Business ethics) forms two network clichés with other topics, two clusters were considered. Ethical cluster #1 is a discussion of ethics in strategy and social business contexts. This cluster appears to be steadily high. On the other hand, ethical cluster #2 (ethics + human rights + supply chain) have declined in popularity. Likewise, environmental cluster displays a slightly declining trend.

5. Discussion

The amount of text data related to CSR is increasing exponentially as traditional reports and documents are digitized, and new digital-based communication channels are available in society. In this vein, text data has become an important source for researching CSR phenomenon. There are diverse sources of data to understand CSR, and traditional communication channels such as corporate reports, news media, and websites are used to understand CSR topics or categories [13,38,39]. Qualitative content analysis has been a popular method for information retrieval for such medium [13,40].

This study has proposed automatic content analysis using topic modeling and social media data to understand what is being discussed about CSR and how different topics are correlated and
have evolved over time. The approach we propose is especially relevant, as digital communication is accelerated among consumers, businesses, and industry associations.

Extant studies and reports of CSR have offered topic categories, including: economic, legal, ethical, and philanthropic [6]; environmental, social, economic, stakeholder, and voluntariness [7]; community focus, consumer orientation, risk management, energy, and health [12]. An industry association (3BL Media’s 100 Best Corporate Citizens 2018) provides seven CSR topic categories, including environment, climate change, human rights, employee relations, corporate governance, philanthropy, and finance. Social media such as Twitter has some uniqueness (e.g., participants, motives, interactions) compared to traditional media [41]. Communication about CSR with Twitter would be less formal and more emotional than with traditional channels, involve an enormous number of industry professionals, businesses, and other stakeholders, and cover content with diverse motives (e.g., promotion, sales). This study provides some patterns that emerge from STM-based topic modeling on a million CSR tweets.

Topic 1 (company strategy) was the most popular topic between 2013 and 2016. This topic exhibits an economic aspect of CSR, the primary responsibility of business [42], suggesting that undertaking CSR is a strategic decision to create a sustainable and competitive corporate value [43]. Another economic topic is Topic 16 (corporate governance), which is also considered a major CSR category from an industry perspective (100 Best Corporate Citizens 2018). Topic correlation shows that Topic 1 is closely associated with other topics (T22: social business, T25: business ethics, T30: entrepreneurship). Three topics in particular (T1, T22, T30) form a network clique showing the popularity surrounding social roles of entrepreneurship and corporate strategy [44]. This seems to concur with the literature discussing such new notions as “sustainable entrepreneurship” and “sustainable entrepreneur” [45,46], where two traditionally separated areas—entrepreneurship and sustainable development—are merged, and new demands are emphasized for entrepreneurs for the triple bottom line.

Topic 15 (community charity) was the second most popular topic overall, and its popularity steadily increased between 2013 and 2016. This topic is mentioned closely with other topics such as Topic 4 (help needed) and Topic 14 (Giving crowdfunding). Topic 15’s high popularity shows that organizations utilize twitter as a tool for promoting community-focused activities and opportunities. Along with Topic 12 (Philanthropy), the prevalence of these topics shows the popularity of community-centric and philanthropic dimensions of CSR [6] on Twitter. CSR is an opportunity to gain organizational legitimacy and reinforce a positive relationship with stakeholders [47]. To maintain corporate legitimacy, it is essential to align corporate activities with social norms, beliefs, and values of stakeholders [48].

While its popularity has decreased over the years, business ethics (T25) ranks 3rd in terms of overall topic prevalence. As mentioned above, the topic is associated with corporate strategy (T1) and social business (T22). In addition, business ethics is discussed with Topic 6 (supply chain) and Topic 20 (human rights), and these three form a network clique. This shows that ethical concerns are expected to be integrated into supply chain operations and management [12]. The correlation between supply chain and human rights shows that issues such as child labor and poor working conditions are big issues in a global supply chain [49].

Topic 18 (climate energy water waste) and Topic 3 (green initiatives) are the next two popular topics, and they are correlated. Analyzing company reports and official websites, the previous study on the firms’ CSR policies [50] also found the firms’ high level of participation in environmental protection to meet the stakeholders’ expectation. Topic 18 is also associated with Topic 6 (supply chain). The literature suggested there are diverse CSR activities related to supply chain [12]. Our study shows that, among those, there is much discussion of energy, climate change, and waste management in the context of sustainable/green supply chain management [51]. These topics represent the environmental dimension of CSR [6].

Some individual topics, not related to any other topics, also deserve discussion. Topic 7 (employee engagement) was almost zero in topic popularity until May 2015 but has since become one of the fastest growing CSR topics in the Twitter-sphere. While employee engagement is not directly found in the popular CSR frameworks [6], it is considered an important strategy for organizations not only to implement their CSR initiatives in scale [52] but also to attract potential employees [53].
Brand marketing (T29) is also shown to be a topic related to CSR in Twitter. The literature has recognized that CSR emerged as a popular strategy for brand management [31] and there are positive effects of CSR on brand performance [54]. Also, the results show that health-related issues (T5) are discussed with CSR. This type of conversation seems to be targeting specific industries such as the tobacco and alcohol industries. However, the health topic was the least popular topic. Some topics such as top latest stories (T23), CSR meetings (T26) and reports and releases (T27) seem to be related to unique characteristics of social media (Twitter in particular) for CSR communication. These topics show that diverse organizations utilize twitter for promotion, information sharing, and announcements [41].

Finally, interest in CSR is growing in developing countries [52]. Two specific countries were popular and captured as a topic: India CSR (T2) and Saudi CSR (T28). These results concur with the literature. For example, Wang, Tong, Takeuchi and George [4] identified increasingly non-US-based studies as one of four major trends in CSR studies. A literature review of CSR studies also shows an increasing number of CSR studies focusing on emerging economies such as India [55–57] and Saudi Arabia [58,59].

6. Conclusions

CSR, along with associated terms such as sustainability, is a popular topic for academic research and business practices. Such academic research and business practices have been generating big data in the form of journal articles, corporate and government reports, news articles, web pages, and social media posts. In response, CSR researchers have attempted to extract valuable information from such text data. Previous studies have tended to focus on traditional mediums such as reports, web pages, and articles using manual or semi-automatic content analysis. Our study has extended this line of research by introducing the potential of structural topic modeling and applying this automatic content analysis method for drawing the topical landscape of CSR from Twitter data. Specifically, this study has aimed to understand what topics are discussed in the Twitter-sphere, how they are associated with each other, and how they have evolved between 2013 and 2016.

Many of our research findings seem to concur with the extant literature. For example, multiple dimensions are found in CSR-related communications in the Twitter-sphere [7,8,42]. Also, these dimensions change over time and trends are formed in discussion CSR [4]. Thus, CSR appears to be an evolving construct in business and society [1]. Our study also reveals that CSR topics tend to be related to each other and these associated topics form topical clusters.

Previous studies suggested that organizations use social media platforms strategically to communicate their efforts with the public [47]. Our study indicates that Twitter is an important medium for organizations to communicate with stakeholders [15,60], especially for promoting, hiring, reporting, and announcing. In this regard, CSR communication in Twitter appears to be different from that of traditional mediums, being less formal and more emotional. This appears to be due to the uniqueness of social media platforms regarding motives, interactions, and participants [41].

Our study has aimed to demonstrate the use of big data and computational techniques in future CSR research and business practices. We suggest there is great potential in big data. To take advantage, CSR researchers and practitioners need to be familiar with computational data collection techniques such as web crawling and APIs. Web crawling is an efficient method of collecting web data, while many big organizations offer APIs through which CSR researchers and practitioners can collect massive amounts of data (e.g., CSR reports, social media posts, news articles). Traditional analysis methods, such as manual or semi-automatic coding, are not suitable for big data. We suggest the use of automatic content methods in CSR research and business practices. Our study has introduced one such method, structural topic modeling (STM). The key advantage of this computational method over traditional methods is the capability of automatically quantifying topic popularity, topic correlation, and topic evolution. This method also allows hypothesis testing based on big text data, although this was not the focus of the present study. Business researchers and practitioners have long endeavored to understand the relationships between independent variables (covariates) and dependent variables.
For CSR practices, given that Twitter is considered a medium to communicate with stakeholders, practitioners are encouraged to establish a sophisticated strategic plan for a CSR communication practice, such as how to interact with their stakeholders in the social media platform. Our study shows that many organizations utilize the social media to promote their CSR practices (e.g., commitment, news, or event), yet it is not clear how extensively they use Twitter to interact with stakeholders proactively. Previous literature emphasized the open discussion or debate about social issues with various individuals to co-create the meaning of CSR [47,61]. When organizations plan to allow open communication with various stakeholders about CSR issues, they should be wary of unintended detrimental impacts that hurt organizations’ reputation.

Two areas—data collection and data analysis—can be improved in future research. First, data collection was done during a limited period (2013 through 2016). Also, the study used a single social media platform as its data source. Future research using decade-long data from multiple data sources would provide a more accurate picture of CSR and its evolutionary pattern in business and society. Second, data analysis relied on a single technique. While topic modeling is an emerging, powerful computational tool for big text data, future research could combine it with other techniques such as sentiment analysis and Twitter user metrics. Sentiment analysis can demonstrate the changes in public sentiment related to CSR topics. Twitter user metrics could be utilized to categorize Twitter users into diverse groups such as individuals, Fortune 500 companies, manufacturing companies, 100 best corporate citizens, etc., and investigate how such groups are using twitter.


Funding: This research received no external funding.

Acknowledgments: Bongsug (Kevin) Chae acknowledges the summer research support from the College of Business at Kansas State University.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A
Figure A1. Cont.
References

3. Aguinis, H.; Glavas, A. What we know and don’t know about corporate social responsibility a review and research agenda. *Bus. Horiz.* 2012, 55, 1–13. [CrossRef]


27. DiMaggio, P. Adapting computational text analysis to social science (and vice versa). *Big Data Soc.* 2015, 2, 2053951715602908. [CrossRef]


38. Capriotti, P.; Moreno, A. Corporate citizenship and public relations: The importance and interactivity of social responsibility issues on corporate websites. *Publ. Relat. Rev.* 2007, 33, 84–91. [CrossRef]


43. Baron, D.P. Private politics, corporate social responsibility, and integrated strategy. J. Econ. Manag. Strategy 2001, 10, 7–45. [CrossRef]


50. Gherghina, S.C.; Vintila, G. Exploring the impact of corporate social responsibility policies on firm value: The case of listed companies in Romania. Econ. Sociol. 2016, 9, 23–42. [CrossRef] [PubMed]


