A COMPUTATIONAL APPROACH TO DETECTING AND ASSESSING SUSTAINABILITY-RELATED COMMUNITIES IN SOCIAL MEDIA

Research-in-Progress

Shan Jiang
Eller College of Management
The University of Arizona
McClelland Hall, Room 430, P.O. Box 210108, Tucson, AZ, 85721-0108
jiangs@email.arizona.edu

Hsinchun Chen
Eller College of Management
The University of Arizona
McClelland Hall, Room 430, P.O. Box 210108, Tucson, AZ, 85721-0108
hchen@eller.arizona.edu

Abstract

The concept of corporate sustainability suggests that firms need to maintain sustainability principles and practices by addressing stakeholders' economic, ecological, and social concerns. Social media has become a knowledge depository where managers can evaluate stakeholders' concerns about the firm's sustainability-related issues. This study proposes a computational approach that utilizes natural language processing techniques to detect sustainability-related communities within online web forums. The validity of the detected communities was assessed based on their impacts on relevant firms' market performance when the firms' social responsibility was challenged. Experiments on three datasets showed that our system is effective in detecting sustainability-related communities. Also, a strong correlation was found between the activities of the identified sustainability-related communities and the firms' market performance during events that challenged the firms' social responsibilities. Our research contributes to the practice of managing corporate sustainability by facilitating managers in evaluating sustainability-related concerns of stakeholders and making effective managerial responses.

Keywords: Corporate sustainability, social media, community detection
Introduction

Sustainable development has been recognized as a critical theme for the society in the last two decades. According to the World Commission on Environment and Development (WCED)’s definition, sustainable development is “development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (Bruntland 1987). The concept has also been adopted by business organizations, referred to as “corporate sustainability” (Linnenluecke and Griffiths 2010; Perrini and Tencati 2006). Corporate sustainability suggests that a firm should not only pursue value maximization for its shareholders, but also focus on long-term value creation in its operations, and maintain sustainability principles and practices by addressing stakeholders’ economic, ecological, and social concerns, known as “triple bottom lines” (Elkington 1997). Failure to addressing these concerns could harm the firm’s long-term success. As a result, many tools and methodologies have been adopted by business organizations for corporate sustainability, such as using the balanced scorecard, integrating sustainability measures into employee evaluations, and providing sustainability reports (SR) to stakeholders. Although these tools contribute to the dissemination of the firms’ efforts, little feedback can be obtained from stakeholders about whether the company’s actions address their concerns and are effective. The critical issues relevant to sustainability vary by company and industry, yet knowing the exact concerns of stakeholders is vital to a firm’s long-term sustainability.

Online social media has become a proxy for business organizations to evaluate stakeholders’ concerns and receive feedback on business activities. Platforms such as firm-specific web forums (e.g. Yahoo! Finance Board), where various stakeholders such as investors and customers discuss the firms’ performance and exchange their opinions, have evolved into a large, dynamic, and real time knowledge depository for research and business practice. These social media platforms can be potentially used by business organizations to understand what sustainability-related issues are of concern to stakeholders and assess their attitudes towards those issues. However, as is typical in these platforms, abundant and noisy information often masks the most important users and sustainability-related content. Consequently, effective approaches are needed to identify the sustainability-related community: the group of users who are most concerned with the company’s sustainability-related issues.

In this study, we developed a computational approach to detect sustainability-related communities in firm-specific web forums. Natural language processing techniques, stylometric analysis and social network analysis were combined to segment the web forum participants into different communities, and the sustainability-related communities were detected based on content analysis within the identified communities. In addition, we used a novel approach to evaluate the detected sustainability-related communities based on their impacts on the firm’s market performance when large unexpected events challenge the firm’s social responsibility.

The remainder of the paper is organized as follows. We first provide a review of relevant literature, based on which potential research gaps are identified. Then we present the framework of our system. The subsequent section describes our testing dataset and experiment settings, and discusses the experiment results. Finally, we conclude the study and suggest potential future work.

Related Works

There are two types of social media communities: explicit and implicit (Papadopoulos et al. 2012). Explicit communities are created consciously by social media users, such as Facebook Groups. Implicit communities are those that naturally emerge from participants’ communications, and the interactions between members of an implicit community are often driven by their shared interests and concerns. Detection of implicit communities in social media has received much attention recently (Choudhury et al. 2010; Gargi et al. 2011; Lin et al. 2009; Papadopoulos et al. 2012). In general, graph-based or clustering-based approaches can be used for implicit community detection. In graph-based approaches, social actors and their relationships (e.g. friendship) are represented by nodes and ties connecting them. Densely connected actors are identified as a community. Examples include the Clique Percolation Method (Palla et al. 2007) and Meta-graph Factorization (Lin et al. 2009). As discovery of a sustainability-related community requires grouping social media users based on their common interests and concerns about a firm, unsupervised clustering-based community detection approaches are more suitable for the task. In
clustering-based approaches, communities are identified based on the similarities between individual members. Various features are first extracted from each individual and the distances between feature vectors are then evaluated to cluster similar individuals into groups. For example, Choudhury et al. (2010) clustered the communication profiles of bloggers and identified “prototypical groups” in blog-spaces. To evaluate the resulting clusters, internal or external criteria can be used. Internal criteria suggest that intra-class similarity should be high while inter-class similarity should be low. In external evaluation, clustering results are evaluated based on external knowledge that is not used for clustering. To evaluate the benefits of detecting sustainability-related communities, external criteria are preferred. However, common external criteria, such as Rand-measure or T-measure, require known class labels of users, which are unavailable for implicit communities. Other criteria that directly address the benefits of detected communities are therefore needed.

Prior studies have suggested several types of features that can be used to characterize each individual social media user. Content-based features characterize each user’s topics of interests by examining the terms and phrases used in the messages posted by the user. To represent content-based features, Bag-Of-Word (BOW) models are typically used (Lee et al. 2010; Schumaker and Chen 2009). Unigrams are often extended to word n-grams to model the phrases. Activity-based features characterize the level of participation of a user and his/her patterns of communication with others. Examples include the number of posts and comments by a blogger (Backstrom et al. 2006), to whom a user replies, and the speed and frequencies of the replies (Choudhury et al. 2010). Stylometry-based features capture the writing styles of a user. The most common stylometry-based feature categories used by prior studies include lexical, syntactic, and structural ones (Abbasi and Chen 2008; Huang et al. 2010; Zheng et al. 2006). Lexical features relate to the character and word usage, such as frequency of letters and vocabulary richness. Syntactic features are about how sentences are constructed by the usage of function words and punctuation. Structural features relate to text organization and layout, such as the number of paragraphs. Prior authorship research has suggested that individuals with similar background and interests tend to have similar writing styles (Baayen et al. 2002). Therefore stylometry-based features may help with grouping social media users with similar interests and concerns about the firm. Although various features have been suggested for community detection, it is unknown that which set of features are best suitable for the discovery of sustainability-related communities in social media.

**System Design**

In this section we describe the framework of the proposed system for detecting and evaluating sustainability-related communities, which is illustrated in Figure 1. Major components of this system are detailed in the subsections.

---

**Figure 1. Framework of Sustainability-related Community Detection System**
**Feature Extraction**

In this step the system extracts three categories of features from each of the users in the web forum to construct feature vectors that represent the user profiles.

For content-based features, proportion of sustainability-related n-grams (n=1, 2, 3) are extracted from all the messages including threads and replies posted by a user. The Encyclopedia of Corporate Social Responsibility is leveraged to identify sustainability-related terms. It provides a list of key terms and phrases that are relevant to corporate sustainability, as well as their definitions and explanations (Idowu et al. 2012). To identify sustainability-related n-grams, stop-words are first removed from the entire list and Porter’s Stemmer is then applied to the remaining terms to unify the inflections (Porter 1980). The resulting terms are then recombined and added to the content-based feature set. Finally, since the terminology in the encyclopedia is generic, the terms with zero occurrences in the dataset are excluded from the feature set. To reduce the computational cost, up to trigrams are used in our system. Table 1 summarizes the content-based features used in our study. The number of features corresponding to each feature subcategory is presented in parentheses.

For activity-based features of each user, the number of threads and replies posted is extracted to account for the user’s level of participation. Also, the number of users communicated with is extracted to model the user’s communication patterns. Table 2 summarizes the activity-based feature set.

For stylometry-based features, the lexical, syntactic, and structural features are extracted from each user. They are summarized in Table 3. The “proportion” in lexical-character (word) features denotes the character (word) count divided by the total number of characters (words). In the lexical feature category, Hapax Legomena and Dislegomena measure the lexical richness of a user (Yule 1944). Other features are straightforward as named.

<table>
<thead>
<tr>
<th>Table 1. Sustainability-related Content Features</th>
<th>Table 2. Activity-based Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
<td><strong>Feature</strong></td>
</tr>
<tr>
<td>Sustainability-related content</td>
<td>unigrams (vary)</td>
</tr>
<tr>
<td></td>
<td>bigrams (vary)</td>
</tr>
<tr>
<td></td>
<td>trigrams (vary)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3. Stylometry-based Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
</tr>
<tr>
<td>Lexical-Character</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Syntactic</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

After the feature extraction phase, each feature is normalized by subtracting its mean and then dividing by the standard deviation. After this step, each user in the web forum is represented by a feature vector \( \mathbf{u'} = (f_1, f_2, ..., f_N) \), where \( N \) is the total number of features.
Community Clustering

Principal Component analysis

As is typical in text processing, feature vectors are characterized by high dimensionality and high correlations among the features, which is problematic for the subsequent clustering phase. To resolve this problem, principal component analysis (PCA) is used to convert the original set of features into a reduced number of principal components (Pearson 1901). Each new component is a linear combination of the original features and the new components are orthogonal to each other. In this study, the number of PCA components is determined as the minimum number with which cumulative variance>0.8. After PCA, original feature vectors are transformed to PCA vectors $v (f_1, f_2, ..., f_N)$, where $V < N$, that represent user profiles.

Expectation Maximization (EM) Clustering

Although various clustering algorithms are available, two aspects of detecting sustainability-related communities guided the selection of the clustering approach used in our study. First, our goal is to detect implicit communities that naturally emerge from users’ activities and discussion content in firm-specific web forums. Although a user may present characteristics that are most evident to a community, the user may also have characteristics that overlap with other community members. Therefore, the assignment of a user to a community should not be dichotomous. Second, the number of implicit communities is unknown before detection. Based on these problem characteristics, the EM clustering method is selected because it assigns instances to clusters probabilistically and automatically determines the optimal number of clusters using cross-validation.

With the EM algorithm, the following generative process of user profile data points is assumed. To generate each user profile $v$:

1) One of the K hidden class labels (i.e., clusters), $z_k$, is picked with probability $p_k$, where $p_1 + p_2 + ... + p_K = 1$.

2) According to the picked class label $z_k$, a user profile is sampled according to a Gaussian distribution: $v|z_k \sim \mathcal{N}(\mu_k, \Sigma_k)$, where $\mu_k$ is the mean vector and $\Sigma_k$ is the variance matrix of the distribution. $\lambda = \{p_k, \mu_k, \Sigma_k | k = 1, 2, ..., K\}$ represents the set of unknown parameters to be estimated.

Under the assumption, the log likelihood of observing a set of data points $V$ is

$$\log L(\lambda) = \log p(V|\lambda) = \sum_{v \in V} \log (\sum_{i=1}^{K} p(v|z_i) \cdot p_i)$$

The EM algorithm initializes $\lambda$ and then repeats the following steps:

1) **E step**: An expected class label for each data point $v$ is calculated based on the following distribution:

$$p(z_k | v, \lambda) = \frac{p(v|z_k) \cdot p_k}{\sum_{j=1}^{K} p(v|z_j) \cdot p_j},$$

The $z_k$ that leads to the highest probability is selected for the class label of $v$.

2) **M step**: Parameter set $\lambda$ is updated to $\lambda'$ by maximizing the lower band of $L(\lambda)$:

$$\lambda' = \text{argmax}_\theta \sum_{v \in V} p(z(v) | v, \lambda) \cdot \log p(v, z(v) | \theta),$$

where $z(v)$ is the expected class label for $v$ from E step. When the EM procedure converges, $z(v)$ is used to assign each data point $v$ to the cluster of corresponding class label. To determine K, we utilized Bayesian Information Criterion (BIC) (Schwartz 1978) to choose the K that minimizes $-2 \log L(\hat{\lambda}) + n \log w$, where $n$ is sample size, $w$ is the number of free parameters in $\lambda$ that results from choosing $K$, and $\hat{\lambda}$ is the estimated parameter values. For more approaches to determine K, readers are referred to (Figueiredo and Jain 2002).
Community Selection

In the last step, the sustainability-related community is selected from the resulting clusters of users based on the keyword proportion (KP) extracted from each cluster. KP measures the proportion of sustainability-related keywords in each cluster. KP for a cluster \( C \) is defined as:

\[
KP(C) = \frac{\sum_{w \in \text{SUS}} \sum_{u \in C} freq(w, u)}{\sum_{u \in C} \text{total # of words by } u}
\]

where \( \text{SUS} \) is the set of non-overlapping sustainability-related n-grams \((n=1, 2, 3)\) derived from Table 1, and \( freq(w, u) \) denotes the frequency of n-gram \( w \) in all the messages posted by user \( u \). Non-overlapping here means that if a lower-order n-gram is part of a higher-order n-gram, the former is excluded from the set \( \text{SUS} \) to avoid recounting it. The higher KP score a cluster has, the higher the chance that the corresponding community is sustainability-related.

Additionally, representative n-grams of each cluster are extracted to indicate what are actually discussed in each identified community. The representativeness score of each n-gram \( w \) in cluster \( C \) is defined in a way similar to TF-IDF in information retrieval:

\[
REP(w, C) = KP(w, C) \cdot \log \frac{\# \text{ of clusters}}{\# \text{ of clusters } C \text{ that satisfies } KP(w, C) \geq KP(w, C)}
\]

where KP\((w, C)\) is the proportion of n-gram \( w \) in cluster \( C \). An n-gram with a high score should have a high proportion in the target cluster to increase KP\((w, C)\), while having a less proportion in other clusters so that the logarithm term in the formula would also take a high value. For each \( n (=1, 2, 3) \), the n-grams with top REP scores are extracted for each cluster.

Assessment of Market Impact

To evaluate the detected sustainability-related community, our system examines how well the activities and message content from this community correlate with the firm’s stock return during events that challenge the firm’s social responsibility. The expectation is that during this period, people would pay more attention to the social responsibility and sustainability-related issues of the firm than usual and relevant discussions would closely relate to the stock returns. Significant correlations between a detected community’s activities and stock returns in this period would suggest the community’s relevance to sustainability and thus support the validity of our proposed system. Similar online community evaluation approaches have been used in prior research (Choudhury et al. 2010).

Previous studies on the relationships between firm-specific web forums and stock return have focused on the number of messages, the average length of messages, sentiment, and disagreement of messages. These web forum metrics, evaluated on a daily basis, do help explain relevant firms’ stock price movements in the same day (Antweiler and Frank 2004; Zhang et al. 2012). Following previous studies, regression models are used to examine the contemporaneous correlations between BP’s stock return and web forum metrics evaluated within sustainability-related communities. Stock return is defined as the log difference of close-to-close stock prices. Message Volume is defined as the number of messages posted per day, logarithmically scaled. Message length is defined as the average word length of the messages each day, logarithmically scaled. The Senti-Word-Net (SWN) Lexicon (Esuli and Sebastiani 2006) is used to calculate the sentiment score of each term and daily sentiment was obtained by averaging these scores. Disagreement is defined as the variance of message-level sentiments per day. Our regression model can be specified as:

\[
STKRET_t = \beta_0 + \beta_1 MSGVOL_t + \beta_2 MSGLEN_t + \beta_3 SENTIMENT_t + \beta_4 DISAGREEMENT_t + \beta_5 MKTRET_t + \epsilon_t
\]

where MKTRET\(_t\) (market return), defined as the log difference of S&P 500 indices, is a control variable that accounts for the overall market condition. The naming of other independent variables and dependent variable is straightforward.
Experiment and Initial Results

Dataset

For an initial test, we used our system to detect sustainability-related communities in British Petroleum (BP), Walmart, and Google. Using an automatic web spidering program, each company’s data was collected from Yahoo! Finance Message Board, where company stakeholders post messages to, for example, discuss various topics, share investment opinions, and express their concerns about the firms. The three companies were selected as test-beds because they are associated with recent major firm-related events that entailed intensive forum discussions regarding social responsibility and corporate sustainability. We focused on the Deepwater Horizon Oil Spill (04/20/10~09/19/10) for BP, a workers’ revolt against new labor policy (10/16/06~03/15/07) for Walmart, and the disclosure of the National Security Administration (NSA)’s mass surveillance (06/06/13~08/06/13) for Google. The selected periods start from the starting days of the events and last for about 2~5 months when intensive discussions about the events took place online. In addition to the postings in event periods, messages were collected for up to one year before the event starting days to allow for forum participants to develop sufficient user profiles. In total we collected 127,246 messages from 11,683 users in BP, 41,512 messages from 3,914 users in Walmart, and 27,517 messages from 2,401 users in Google.

Detection of Sustainability-related Communities

Community clustering was implemented on each of the three companies using the entire data collection. After PCA, the reduced feature dimensions were 40 for BP, 51 for Walmart, and 37 for Google. Table 4 shows the proportion of users, KP, and representative n-grams of the clusters resulting from our experiment on BP’s forum and partial results in Walmart and Google’s forums. Clusters with less than 1% of users were omitted from the table. Representative unigrams were also excluded because they tended to be general terms and were less indicative of cluster contents compared to bigrams and trigrams.

For BP, Cluster B2 received the highest KP score, followed by cluster B5. Based on representative n-grams, these communities were mainly concerned about environmental and labor issues, respectively. Other clusters had either much lower KP scores, uninformative representative n-grams, or both. As a result, clusters B2 and B5 were selected as the sustainability-related communities in BP, which altogether accounted for about one-fourth of all users in BP’s board. With the same approach, three sustainability-related communities from Walmart’s board, and two sustainability-related communities from Google’s board were detected, as shown in the table. Due to the exploratory nature of this study, the above community selection procedure involved human judgment and was not completely quantitative. However, based on our observation, we suggest using KP>3% as the selection criteria.

Table 4. Identified Clusters in BP and Sustainability-related Clusters in Walmart and Google

<table>
<thead>
<tr>
<th>Forum</th>
<th>Cluster</th>
<th>Users</th>
<th>KP</th>
<th>Selected Representative N-grams (N=2, 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>B1</td>
<td>51.0%</td>
<td>.0137</td>
<td>stock dividend, cash flow, buy back, stock price, money supply</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>24.4%</td>
<td>.0614</td>
<td>global warm, shark extinct, environment protect, clean air act</td>
</tr>
<tr>
<td></td>
<td>B3</td>
<td>11.2%</td>
<td>.0002</td>
<td>lol lol, click link, http www, ha ha ha</td>
</tr>
<tr>
<td></td>
<td>B4</td>
<td>5.1%</td>
<td>.0171</td>
<td>crude oil, oil inventory, million barrel day</td>
</tr>
<tr>
<td></td>
<td>B5</td>
<td>4.9%</td>
<td>.0321</td>
<td>increase wage, sexual orientation, equal employ. opportunity</td>
</tr>
<tr>
<td>Walmart</td>
<td>W2</td>
<td>15.8%</td>
<td>.0874</td>
<td>child labor, poor condition, equal employ. opportunity</td>
</tr>
<tr>
<td></td>
<td>W3</td>
<td>7.6%</td>
<td>.0613</td>
<td>parent control, wild life, wise business decision</td>
</tr>
<tr>
<td></td>
<td>W6</td>
<td>11.0%</td>
<td>.0378</td>
<td>property damage, grocery store, civil war battlefield</td>
</tr>
<tr>
<td>Google</td>
<td>G2</td>
<td>29.8%</td>
<td>.0598</td>
<td>privacy policy, big data, national security agency</td>
</tr>
<tr>
<td></td>
<td>G4</td>
<td>14.4%</td>
<td>.0319</td>
<td>google plus, google play, click through rate</td>
</tr>
</tbody>
</table>

Evaluation

To evaluate the detected sustainability-related communities, we examined how well the activities of these communities correlated with the relevant companies’ stock return during the events that challenged corporate sustainability, as mentioned before. Table 5 shows the regression coefficients, p-values (shown in parentheses), R-squared and mean squared errors (MSE) evaluated within various communities of BP,
Walmart, and Google during the respective event periods. For baseline model, we used the web forum metrics evaluated by all the users in each company’s forum. The sustainability-related communities in BP are B2 and B5 as discussed before. Due to space limitations, we only show selected sustainability-related communities that outperformed the baseline models in Walmart (W3) and Google (G2) in the table. The $R^2$ of all models (0.082–0.112) were comparable to that of previous finance research that examined the relationships between web forums and stock returns (Antweiler and Frank 2004; Zhang et al. 2012). In BP's forum, we found that the daily sentiment of B2, the community with concerns about environmental issues, had significant and positive correlation with BP's stock return. This relationship was not observed in BP's baseline model. Also, B2 outperformed the baseline in terms of higher $R^2$ and lower MSE. This suggests that the detected sustainability-related community B2 had stronger relationships with BP’s stock return during the oil spill period, compared to the entire forum. Interestingly, when we ran the same regression using data five months before the oil spill, such strengthened relationships between B2 and stock return were not observed, implying that the correlation between the sustainability-related community B2 and BP’s stock return only manifested during the event period. The other sustainability-related community B5 did not show a strengthened relationship with BP’s stock return during the oil spill period. Nevertheless, it may do so during other types of events. Therefore, in order to maintain corporate sustainability, a firm should pay attention to all sustainability-related communities and not neglect any. Similarly, we also observed that the communities W3 in Walmart's forum and the G2 in Google's forum outperformed their baseline models in terms of higher $R^2$, and lower MSE during their respective event periods. Different from their baseline models, the daily average message length of W3 and the sentiment of G2 had significant correlations with respective firm's stock return during event periods. Overall, the results in Table 5 suggest that our system is effective in detecting sustainability-related communities that are strongly correlated with firms' stock return during major firm events.

| Table 5. Regression Coefficients, R-squared, and P-values During Event Periods |
|---------------------------------|---------|---------|---------|---------|---------|---------|
|                                 | BP      | Walmart | Google  |
| **Independent Variables**       | **Baseline** (all users) | **Sus. B2** | **Sus. B5** | **Baseline** (all users) | **Sus. W3** | **Baseline** (all users) | **Sus. G2** |
| MSGVOL                          | .007* (.047) | .010* (.036) |.004* (.084) | .003* (.063) | .006* (.052) | .016* (.056) | .011* (.050) |
| MSGLEN                          | -.001 (.286) | -.014 (.158) |-.105 (.351) | .001 (.178) | -.002* (.051) | -.004 (.565) | -.004 (.429) |
| SENTIMENT                       | .005 (.374) | .168* (.014) |-.014 (.224) | .010 (.426) | .013 (.378) | .016 (.322) | .085* (.094) |
| DISAGREEMENT                    | .000 (.893) | .004 (.779) |-.001 (.513) | .000 (.832) | .000 (.629) | -.001 (.792) | -.0012 (.352) |
| MKTRET                          | .034** (.000) | .032** (.000) |.034* (.000) | .042** (.000) | .040** (.000) | .031** (.000) | .033** (.000) |
| $R^2$                           | .092 | .100 | .082 | .112 | .121 | .086 | .090 |
| MSE                             | .01436 | .01414 | .01482 | .02883 | .02816 | .01725 | .01698 |

*p<0.1, **p<0.01

Furthermore, in order to find the best feature set for the sustainability-related community detection task, we tried using different combinations of the three feature categories 1) only content-based features, 2) content+ activity, 3) content + stylometry, and 4) content + activity + stylometry to adjust B2, W3, and G2, and ran regression models during the event periods. Figure 2 compares the $R^2$ and MSE performance of each model in BP’s forum. Among the four sustainability-related community models, model 4, which
utilizes all features, resulted in the best performance. The performance of model 4 was significantly better than model 1 in terms of higher $R^2$ and lower MSE, indicating that the addition of activity and stylometry features improved the effectiveness of sustainability-related community detection, compared to the approach solely relying on sustainability-related terms. However, the improvements from model 1 to model 2 and from model 3 to model 4 were less obvious, suggesting that the activity-based features added marginal benefits. One explanation is that the activity features did not differ greatly between sustainability-related communities and others. For example, the percentage difference of the number of threads per user between sustainability-related community and the rest in BP was only 8.4%, and the percentage differences of other activity-based features were all below 20%. Experiments on Walmart and Google’s data showed the same qualitative results (content + activity + stylometry had the best performance; benefits of incorporating activity-based features were marginal), supporting the above findings. Due to space limitations, we omit the quantitative results for the two companies.

![Figure 2. Performance Comparison between Four Sustainability-related Models in BP's Forum](image)

**Conclusion**

In this study, we proposed a novel computational approach to detect and evaluate sustainability-related communities in firm-specific web forums. The following initial observations were made based on our experiments. First, our system was effective in detecting sustainability-related communities and extracting their concerns related to the firm. Although such communities may not have been the largest in size, their opinions and concerns about the firm could be crucial for the firm’s long-term success. Second, incorporating activity and stylometry information of users led to a more effective way of detecting sustainability-related communities than the detection approach based solely on sustainability-related terms and phrases. Finally, during the period of time when major events challenge firms’ social responsibility, activities and opinions of some detected communities may have great impacts on the firms’ stock returns, indicating that firms should pay attention to all its sustainability-related communities, who are of great value to long-term corporate sustainability management.

The contributions of this study are two-fold. First, our research contributes to the practice of managing corporate sustainability. Based on our system, managers of a firm can evaluate and address sustainability-related concerns in real time, and take effective managerial responses when important events and sustainability challenges occur. Second, from the methodology perspective, the market impact assessment used in this study acts as an example of utilizing external information to evaluate clustering results. This approach might be useful in clustering evaluation in a business context. In future work, we will include multiple types of events to examine how different sustainability-related communities detected from the same company react, which was not explored fully in this study. We will also test our system on a wider variety of companies and industries to validate its generalizability.

**Acknowledgments**

The research is based upon work supported in part by the National Science Foundation under Grant No. CMMI-1057624. The authors are thankful to to Cathy Larson for her suggestions and comments.
References


Press.