COLOR REPRODUCTION IN YOUR ARTICLE

These proofs have been typeset using the original figure files transmitted to production when this article was accepted for publication. Please review and mark your approval of each figure individually within your proof corrections. Should you need further assistance, please contact journal production at jasistprod@wiley.com

Because of the high cost of color printing, we can only print figures in color if authors choose to cover the expense. If you have submitted color figures, please indicate your consent to cover the cost on the table listed below by marking the box corresponding to the approved cost on the table. The charge for printing figures in color is $800 per figure.

Please note, all color images will be reproduced online at no charge, whether or not you opt for color printing.

You will be invoiced for color charges once the article has been published in print.

Failure to return this form with your article proofs will delay the publication of your article.

JOURNAL Journal of the Association for Information Science and Technology

MS. NO. ___________ NO. COLOR FIGURES ___________

MANUSCRIPT TITLE ___________________________

AUTHOR(S) ___________________________

<table>
<thead>
<tr>
<th>No. Color Figures</th>
<th>Color Charge</th>
<th>No. Color Figures</th>
<th>Color Charge</th>
<th>No. Color Figures</th>
<th>Color Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$800</td>
<td>5</td>
<td>$4000</td>
<td>9</td>
<td>$7200</td>
</tr>
<tr>
<td>2</td>
<td>$1600</td>
<td>6</td>
<td>$4800</td>
<td>10</td>
<td>$8000</td>
</tr>
<tr>
<td>3</td>
<td>$2400</td>
<td>7</td>
<td>$5600</td>
<td>11</td>
<td>$8800</td>
</tr>
<tr>
<td>4</td>
<td>$3200</td>
<td>8</td>
<td>$6400</td>
<td>12</td>
<td>$9600</td>
</tr>
</tbody>
</table>

***Contact jasistprod@wiley.com for a quote if you have more than 12 figures in color***

☐ Please print my figures in color
☐ Please print my figures in black and white

☐ Please print the following figures in color

and convert these figures to black and white

Approved by ___________________________

Billing Address ___________________________ E-mail ___________________________

Telephone ___________________________ Fax ___________________________
Dear Author,
During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof.
Many thanks for your assistance.

<table>
<thead>
<tr>
<th>Query References</th>
<th>Query</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Please confirm that given names (blue) and surnames/family names (vermilion) have been identified and spelled correctly.</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>Please check if link to ORCID is correct.</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>Note that abstract is missing in the source doc, hence abstract has been taken from meta. Please check and confirm the insertion of abstract.</td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td>Reference “Zheng, 2014” has been changed as “Zheng et al., 2014” to match reference list. Please confirm whether it is OK as set.</td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td>Please provide unlisted reference “Alexa, 2012.”</td>
<td></td>
</tr>
</tbody>
</table>
Public Health and Social Media: A Study of Zika Virus-Related Posts on Yahoo! Answers

Jin Zhang  
School of Information Studies, University of Wisconsin Milwaukee, Milwaukee, WI. E-mail: jzhang@uwm.edu

Ye Chen  
School of Information Management, Central China Normal University, Wuhan, China. E-mail: cheny@mail.ccnu.edu.cn

Yuehua Zhao  
School of Information Management, Nanjing University, Nanjing, China. E-mail: yuehua@nju.edu.cn

Dietmar Wolfram  
School of Information Studies, University of Wisconsin Milwaukee, Milwaukee, WI. E-mail: dwolfram@uwm.edu

Feicheng Ma  
School of Information Management, Wuhan University, Wuhan, China. E-mail: fchma@whu.edu.cn

This study investigates the content of questions and responses about the Zika virus on Yahoo! Answers as a recent example of how public concerns regarding an international health issue are reflected in social media. We investigate the contents of posts about the Zika virus on Yahoo! Answers, identify and reveal subject patterns about the Zika virus, and analyze the temporal changes of the revealed subject topics over four defined periods of the Zika virus outbreak. Multidimensional scaling analysis, temporal analysis, and inferential statistical analysis approaches were used in the study. A resulting two-layer Zika virus schema, and term connections and relationships are presented. The results indicate that consumers’ concerns changed over the four defined periods. Consumers paid more attention to the basic information about the Zika virus, and the prevention and protection from the Zika virus at the beginning of the outbreak of the Zika virus. During the later periods, consumers became more interested in the role that the government and health organizations played in the public health emergency.

Introduction

Social question and answer (Q&A) websites provide health consumers with opportunities to discover information online (Morris, Teevan, & Panovich, 2010). A social Q&A website such as Yahoo! Answers allows users to make contributions to a topic of interest, receive assistance from others, and interact with each other in a community. Health consumers rely more and more on Q&A websites to find health information to solve their health problems. The wealth of online information found on social Q&A websites has been the center of much research. Research on social Q&A websites can be classified into user-centered studies (questions, answers, and the community in general) (Wu & He, 2014) and content-centered studies (questions and answers) (Oh, 2018; Kim & Oh, 2009). The major perspectives of social Q&A website research include the motivation for participation (Jin, Li, Zhong, & Zhai, 2015; Zheng et al., 2014), user satisfaction (Zhang, Deng, & Yang, 2014), user behavior (Park, Reber, & Chon, 2016; Shoji, Fujita, Tajima, & Tanaka, 2015; So et al., 2016), and information quality and effectiveness (Chua & Banerjee, 2015; Li, He, Jeng, Goodwin, & Zhang, 2015; Robillard, Johnson, Hennessey, Beattie, & Illes, 2013). In addition, user information needs analysis has become a popular research interest using social Q&A websites.

The Zika virus serves as an example of a recent, international disease that has attracted public attention on social media. The Zika virus was first identified in Uganda in 1947 in monkeys. It was later found in humans in 1952 in Africa. It gained prominence due to the increase in cases in Brazil...
and the spotlight on the 2016 Summer Olympic Games in Rio de Janeiro. The Zika virus, normally spread by mosquitoes, can also be transmitted through sexual contact and blood transfusions (Hennesy, Fischer, & Staples, 2016). It can be passed from a pregnant woman to her newborn baby, causing severe brain malformations, birth defects, and other symptoms. Due to its transmission characteristics, the Zika virus is considered a severe infectious disease. In addition, there are no effective modern medications or vaccines to prevent or cure it. It is not surprising that the outbreak of the Zika virus can lead to public panic, and health consumers use social media like a public Q&A forum to seek related information to address their concerns about the disease. As a result, an investigation of the Zika virus outbreak and health consumers’ responses to it on Yahoo! Answers, is warranted.

The primary aim of this study was to investigate Zika virus-related posts on Yahoo! Answers. Toward this end, the following objectives are presented: (i) we investigated the contents of posts about the Zika virus on Yahoo! Answers; (ii) we identified and revealed subject patterns from the posts; and (iii) we analyzed the temporal changes of the revealed subject topics during the Zika virus outbreak.

This study has both theoretical and practical implications. The results of the study can help healthcare providers/professionals, health stakeholders, and health policy makers better understand the health consumers’ concerns about the virus. The revealed consumer-oriented schema can be used to organize information on the Zika virus-related websites, digital libraries, information repositories, Internet search engines, and databases. The discovered terms relationships and patterns can be employed to enrich existing Zika virus-related medical thesauri, subject headings, and classification systems. Term connections can be utilized to enhance search engines and, finally, the Zika virus schema can be used for sophisticated semantic searches. The method used in the present study may be applied to other diseases and may serve as a model for studies of similar diseases that attract broad public attention.

Related Research Studies

Due to the richness and availability of health information on the Internet, people like to search online for this information. Studies on the social life of health information seekers reported that 72% of adult Internet users have searched online for health-related information (Fox, 2014).

When concentrating on health information needs analysis on social Q&A websites, the core research questions of health information needs analysis have included the concerns of health information consumers and the characteristics of their health information needs. All health-related questions/answers usually have been regarded as health information needs. Both general health topics (Lu, Lu, Jeng, Farzan, & Lin, 2015; Oh, Lauckner, Boehmer, Fewsins-Bliss, & Li, 2013) and specific diseases such as cancer, diabetes, eating disorders, sexually transmitted diseases (STDs), and smoking cessation (Bowler, Mattern, Jeng, Oh, & He, 2013; Bowler, Oh, He, Mattern, & Jeng, 2012; Oh, Yan, & Min, 2012; Ramo, Liu, & Prochaska, 2015; Tsuya, Sugawara, Tanaka, & Narimatsu, 2014; Zhang & Zhao, 2013; Zhang et al., 2014), have attracted the attention of researchers. In addition, a framework for modeling health information needs containing types of information and a taxonomy for underlying user intent containing have been developed (Chen, Zhang, & Mark, 2012, April; Oh, 2015; Oh et al., 2012). There are many public Q&A platforms like Stack Exchange, Quora, HealthTap, and Patientslikeme. We selected Yahoo! Answers for this study because it was considered to be one of the most important public Q&A platforms in terms of its coverage, public visibility, and quality (Srba & Bielikova, 2016; Alexa, 2012).

Temporal analysis has been considered a valuable means to analyze activities beyond simply stating facts, to ascertain both contextual and temporal relationships, and finally to assist in understanding the past and forecasting the future (Yang & Yang, 2015). However, only a limited number of articles have applied a temporal analysis method to the healthcare domain. For instance, it was employed to identify the evolutionary stages of melanoma through online melanoma forums (Durant, McCray, & Safran, 2010); to quantify the responsiveness of interactions on online medical forums (Durant, McCray, & Safran, 2011); and to detect drug safety signals using health-consumer-contributed data (Yang & Yang, 2015).

Relatively few research studies have been conducted on health information needs related to the Zika virus on social media. Content-based analysis has been applied to tweets to unveil the public concerns about the Zika virus. Fu et al. (2016) analyzed the contents of 62,547 Zika-related tweets and revealed five themes using topic modeling. Glowacki, Lazard, Wilcox, Mackert, and Bernhardt (2016) collected tweets from a twitter live chat hosted by the Centers for Disease Control in the United States and summarized 10 themes for public concerns related to the Zika virus. Likewise, Miller, Banerjee, Muppalla, Romine, and Sheth (2017) extracted detailed information from discussions on Twitter and identified 20 subtopics associated with Zika symptoms, transmission, prevention, and treatment. Using Facebook as an information resource, Sharma, Yadav, Yadav, and Ferdinand (2016) discovered that misguided video posts about the Zika virus were far more popular than the posts disseminating accurate information about the Zika virus. Bragazzi et al. (2017) conducted a study based on data from Twitter, Google Trends (GT), Google News, YouTube, and Wikipedia search queries and their findings demonstrated that the main spikes of public interest in the Zika virus occurred first in February 2016 and a second in August 2016. Beyond these studies, there has been little research on Zika-related contents on social Q&A websites.

We investigated consumers’ comments (questions and answers) about the Zika virus on Yahoo! Answers, which reflect their concerns about the virus, and explore changes in consumers’ information needs over time during the Zika virus outbreak and we develop a subject schema related to the Zika virus.
Research Method

Selection of a Social Q&A Website

We selected the social Q&A website Yahoo! Answers (2017) because it is regarded as one of the most important and widely used public social Q&A websites (Srba & Bielikova, 2016). It includes more than 21 million unique users in the U.S. and 90 million users worldwide. Yahoo! Answers also provides a subject directory containing 26 classes.

Data Collection

An exhaustive search on Yahoo! Answers was conducted to include all items relevant to the Zika virus. The search queries included “the Zika virus,” “the Zika disease,” “the Zika virus disease,” “Zika,” “Zika disease,” “Zika virus,” and “Zika virus disease.” The search was expanded beyond the health category. In a pilot study, we found that the first item relevant to the Zika virus on Yahoo! Answers was posted on December 25, 2015. The last Zika outbreak happened in 2015–2016. Therefore, the posts for this study were collected from December 25, 2015 to January 1, 2017. After all data were collected, an ID was assigned to uniquely identify each record. Next, the items were manually examined to determine their inclusion in the study. Items which were not relevant to the Zika virus were removed. As a result, 209 records were collected. For each Zika virus-related record/item the content of the question/answer and the time stamp were extracted.

We used the software application Bazhuayu (2013) for data collection.

Term Cleansing and Normalization

The raw data collected from Yahoo! Answers were manually cleansed and normalized. First, all words were extracted from a record/item to produce a word list. All words on the list were tallied to calculate a frequency for each word. Second, stopwords were removed from the list because these words carried no meaning without context after they were extracted from the texts. Third, the words were normalized by converting misspelled words, converting irregular words to their regular forms, and changing inflectional and derivational variants of words to their basic lexical forms. Fourth, synonyms were merged. Each of the extracted words was examined against the database Wordnet (Miller, 1995) to find its synonyms, and all synonyms were merged into a regular word.

Next, the normalized terms with low frequencies were removed from the list because these terms made little contribution to the term visualization/clustering analyses. Terms whose frequencies were equal to 1 did not co-occur with other words and could not be grouped with other words in the clustering analysis. These words were eliminated from the normalized term set. Other low-frequency words also did not contribute significantly to the clustering analysis. Consequently, the term frequency cutoff thresholds in this study were set from 2 to 4. After the term cleansing and normalization, each Q&A record/item corresponded to a set of the normalized terms.

Q&A Item Classification

The primary themes/categories were identified and produced based on the thematic analysis method. The following steps were used for the thematic analysis: collection and familiarization with the data, generation of initial codes, identification of themes among codes, review of themes, definition and naming of themes, and production of the final results (Braun & Clarke, 2008). The pilot study revealed four categories (About the Zika virus, Self-expression, Related event, and others). Each of the items was examined, its subject was analyzed and reviewed, and then it was classified into a category in a mutually exclusive way. The category titles were assigned and adjusted by the researchers during the process. The final analysis resulted in three revised categories. Each of the collected items was manually examined and its content was analyzed independently by two of the researchers (Y.C. and Y.Z.). Consequently, three primary themes/categories were discovered. They were: About the Zika virus, Social impact of the Zika virus, and Consumer concerns about the Zika virus. A Cohen’s kappa intercoder reliability value was calculated on a random sample of the data set. The resulting value of k = .8 indicated a high level of intercoder agreement. Identification of these primary categories enabled the researchers to analyze the changes of these categories in a temporal analysis, and to lay a foundation for a more detailed clustering analysis within each of these primary categories using a visualization method.

Temporal Analysis

Temporal analysis refers to the process of recording a series of observations obtained from objects over time, and then analyzing the change of the objects over defined periods. It usually involves repeated observations of the same objects in the defined periods. As a result, the impact of time on the objects can be discovered and compared.

The outbreak and effective control of the Zika virus happened during a certain time period. The outbreak of the Zika virus led to an immediate increase in the number of questions and comments about the virus on Yahoo! Answers. The number of questions and answers about the virus on Yahoo! Answers declined dramatically after the virus was effectively controlled. Therefore, time is an important variable to analyze questions and answers of public health consumers about the Zika virus from the outbreak to the effective control of the disease.

The first question about the Zika virus was posted on Yahoo! Answers on December 25, 2015. In February 2016, the World Health Organization (WHO) declared the Zika virus a Public Health Emergency of International Concern due to an outbreak that started in Brazil in 2015 (WHO, 2016b). In November 2016, WHO lifted its commitment to a long-term response to the Zika virus as “public health emergency” (WHO, 2016a). In this study, the entire period...
of the data collection was from December 25, 2015 to January 1, 2017. The data collection covered the period of the outbreak of the Zika virus defined by WHO. Notice that if the number of the periods is too small, it is difficult to discover the pattern of the post changes over the periods in the temporal analysis. If the number of the periods is too large, some periods may receive few posts and it would be hard to conduct a visualization analysis for each of the periods because the terms cannot effectively be visualized and clustered in the visual space. In addition, the entire period should cover the major events related to the outbreak of the Zika virus. The first question and answer item on the Zika virus on Yahoo! Answers was posted on December 25, 2015, and the number of Zika virus-related posts decreased to zero in December, 2016 and January, 2017 as well. As a result, in the study the number of the periods was defined as four. Each period was approximately 3 months. Finally, the entire period was divided into four periods: Period I: December 25, 2015 to March 31, 2016; Period II: From April 1 to June 30, 2016; Period III: From July 1 to September 30, 2016; and Period IV: October 1, 2016 to January 1, 2017.

Data Organization and Definition

An item/record on a social Q&A website consists of a question and the corresponding answers posted by users. The item, an independent logic data unit, contains a data set of attributes rich in information. To analyze and reveal hidden themes from the data set, these items were organized in a vector space. They can be organized in an item/attribute matrix. In the matrix, the rows are the items while the columns are the attributes or terms in this case.

\[
M_{IA} = \begin{pmatrix} t_{11} & \ldots & t_{1n} \\
\vdots & \ddots & \vdots \\
t_{k1} & \ldots & t_{kn} \end{pmatrix}
\]  

(1)

\[M_{IA}\] is the item/attribute matrix. Here \(n\) is equal to the number of the items in the data set while \(k\) is equal to the number of the attributes in the data set. \(M_{IA}\) is a \(n \times k\) matrix.

An attribute/attribute matrix \(M_{AA}\) was generated from the following equation:

\[
M_{AA} = M_{IA} \otimes M_{IA}^T
\]

(2)

In Equation (2), \(\otimes\) denotes the Cosine similarity measure and \(M_{IA}^T\) refers to a transposed matrix of \(M_{IA}\). After this process, \(M_{AA}\) became a \(k \times k\) matrix.

The attributes were the normalized terms extracted from the questions and answers in the data collection process. The \(k \times k\) matrix \((M_{IA})\) served as an input data set for the Multidimensional Scaling (MDS) analysis.

MDS Analysis

MDS analysis reduces the high dimensionality of the information space and projects the items from a high dimensional space onto a low dimensional space (a two-dimensional or three-dimensional space) where the semantic relationships among the items can be preserved and observed (Zhang, 2007).

In the MDS analysis process, the dimensionality of the high space was reduced, and the terms were projected onto a low dimensional space for observation. The low space was defined in a Euclidean space. The strength between two terms in the high dimensional space was calculated by the Minkowski distance. The distance was then assigned to each axis of the low space. In this case the low space was three-dimensional; the distance was assigned to the X-axis, Y-axis, and Z-axis, respectively.

The relationships among the terms in the high dimensional space should be faithfully represented in their relationships in the low dimensional visual space after the projection. Toward this aim the difference between the strengths among the terms in the high dimensional space and their strengths in the low dimensional space was monitored during the MDS process. The positions of the projected terms in the low space were constantly adjusted so that the difference was minimized to achieve the best match between the terms in the two spaces during the MDS process.

After an MDS process was completed, all terms were projected onto the visual space for further analysis. The terms that were semantically related were clustered and grouped together in the visual space. In order to easily identify the terms, the researchers used circles to specify different resultant clusters in the visual space.

MDS was used because it has the following advantages: the data are relatively free of any distributional assumptions; MDS can effectively process various types of data like ordinal, interval, and ratio-level data; it presents an intuitive visual display for the clustering analysis results; it is mature and widely used in many application domains; and it offers an interactive environment in the SPSS software (IBM, Armonk, NY) which allows the researchers to observe term relationships from any angle, thus avoiding possible ambiguity. These were crucial for the researchers to identify the meaningful term clusters in the visual space, and then utilize various similarity measurements to produce a sound visual display for analysis. MDS was used to identify clusters/categories, reveal hidden themes from the collected data, and form the Zika virus-related schema.

SPSS (v. 25) allows the researchers to observe the MDS display of the terms from an angle at will in a 3D space. As a result, researchers can select a proper observation angle to observe the relationships of the projected terms clearly, avoid the unnecessary overlapping terms, which can lead to ambiguity in the term relationships, and group the terms accurately in the space.

The INDSCAL model in SPSS was used for the MDS analysis. The quality of results for an MDS analysis is usually measured by two indicators: stress value and the coefficient of determination (R²). Generally speaking, the lower the stress value (or the higher the R² value), the...
better the MDS result is. In other words, a low stress value (or a high $R^2$ value) corresponds to a plausible and reliable MDS analysis result.

**Inferential Analysis**

The following null hypothesis was introduced to ascertain the changes of the revealed themes over the defined periods:

$H_0$: There are no significant differences between the revealed theme categories and the defined four periods in terms of the questions and answers about the Zika virus on Yahoo! Answers.

The significance level for this test was set to 0.05.

**Results and Analysis**

**Summary of the Collected Data**

Question and answer data were collected from Yahoo! Answers. A total of 209 items related to the Zika virus were retrieved and validated from Yahoo! Answers. Each item contained a question and the corresponding answers. *Politics & Government, Health, Society & Culture, Entertainment & Music, News & Events, Travel, Science & Mathematics, Pregnancy & Parenting, Sports, Education & Reference, Family & Relationships, Dining Out, Environment, Social Science* received 75, 34, 31, 14, 13, 12, 10, 7, 6, 2, 2, 1, 1, and 1 items, respectively. They covered 14 Yahoo! Answers classes, which represented more than half of its total classes (26). *Politics & Government* (75) outnumbered *Health* (34) in terms of the posts, and *Society & Culture* (31) followed *Health* (34). It is not surprising that the impact of the Zika virus outbreak went far beyond people’s health, and it extended to government policy, social and cultural events, everyday activities, and entertainment.

The relevant items related to the Zika virus that were posted between December 25, 2015 and January 1, 2017 and their frequencies are illustrated in Table 1. There were two obvious spikes, in February 2016 (39) and August 2016 (58). The entire investigated period was divided into four periods as described in Temporal Analysis section. Period III (92) had more posts than the other periods and period IV received the lowest number (17). The average number of the items per period was 52.5.

**Visual Theme Analysis**

Three primary themes/categories arose from the content analysis for the pilot study: About the Zika virus (Category A), Social impact of the Zika virus (Category B), and Consumer concerns about the Zika virus (Category C).

To ensure the quality of the coding results, an intercoder consistency test was conducted. To test the intercoder reliability of assigning categories, two researchers independently coded 100 items (47.8% of the entire data set) randomly selected from the data set. The kappa reliability analysis method (Cohen, 1960) was applied to the results. The resultant intercoder reliability was 0.80. This means that the two coders achieved a substantial agreement in the category assignments.

Visualization analyses were conducted for each of the four periods.

**Period I** (12/25/2015–03/31/2016). The visual analysis results for Categories A, B, and C in Period I are displayed in Figures 1–3, respectively. The terms within each subcategory are listed in Table 2. The measurement, stress value, and $R^2$ value for each category for each time period appear in Table 6.

**Period II** (04/01/2016–06/30/2016). The visual analysis results for Categories A, B, and C in Period II are displayed in Figures 4–6, respectively. The terms within each subcategory are listed in Table 3.

**Period III** (07/01/2016–09/30/2016). The visual analysis results for Categories A, B, and C in Period III are displayed in Figures 7–9, respectively. The terms within each subcategory are listed in Table 4.

**Period IV** (10/01/2016–01/01/2017). The visual analysis results for Categories A, B, and C in Period IV are displayed in Figures 10–12, respectively. The terms within each subcategory are listed in Table 5.

All results of the MDS analyses are listed in Table 6. The stress values are each below 0.20 and the $R^2$ values above 0.80, indicating that the results are sound and acceptable.

The schema labels were defined and assigned by the researchers. The labels were defined in a general way to be inclusive. Labeling a subcategory was based on both the terms in a cluster in the visualization analysis and their contexts in

**TABLE 1.** Descriptive statistics of the four periods.

<table>
<thead>
<tr>
<th></th>
<th>Period I (12/25/15–03/31/16)</th>
<th>Period II (4/1/16–06/30/16)</th>
<th>Period III (7/1/16–09/30/16)</th>
<th>Period IV (10/1/16–01/01/17)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of questions</td>
<td>64</td>
<td>36</td>
<td>92</td>
<td>17</td>
<td>209</td>
</tr>
<tr>
<td>Total number of answers</td>
<td>349</td>
<td>162</td>
<td>381</td>
<td>42</td>
<td>1237</td>
</tr>
<tr>
<td>Average number of answers</td>
<td>5.5</td>
<td>4.5</td>
<td>4.1</td>
<td>2.5</td>
<td>5.9</td>
</tr>
<tr>
<td>Average number of words per question</td>
<td>59.4</td>
<td>57.5</td>
<td>66.4</td>
<td>89.5</td>
<td>68.2</td>
</tr>
<tr>
<td>Average number of words per answer</td>
<td>73.6</td>
<td>85.1</td>
<td>55.4</td>
<td>56.2</td>
<td>67.6</td>
</tr>
<tr>
<td>Average number of words per item</td>
<td>460.8</td>
<td>440.5</td>
<td>295.8</td>
<td>228.3</td>
<td>356.4</td>
</tr>
</tbody>
</table>
the corresponding question and answer records. For instance, the terms Christmas, holiday, and avoid appeared in a cluster of a visualization display, and the cluster was labeled as Prevention and solution in Table 3 because the terms were found in records like “How do I avoid getting Zika this Christmas? I don’t want to miss the holiday because of it.” The theme of the record was about prevention.

Resultant schema. Based on the previous content analysis and visualization analysis results, the categories and sub-categories were generated for a Zika virus schema. A two-layer schema of consumer health information needs about the Zika virus is displayed in Figure 13. The first layer contains three earlier defined categories. The second layer consists of seven subcategories. The first category contains causes and symptoms, and transmission and risks. The second category contains protection and surveillance; activities and responses, and social events and travel. Lastly, the third category contains attitudes and opinions, prevention and solutions.

Temporal Analysis

A chi-square test for independence was conducted to examine the topic variation across the four periods for the questions and answers. The results for the questions are: $df = 6, N = 209$, $\chi^2 = 14.2$, $p = .02$. The effect size Cramer’s $V$ is 0.184, which is considered a medium effect. The results for answers are: $df = 6, N = 1237$, $\chi^2 = 101.9$, $p = .00$. The effect size Cramer’s $V$ is 0.203, which is considered a medium effect. Since the two $p$-values were smaller than the significance level (.05), both hypotheses were rejected. That is, there were significant differences between the revealed theme categories and the four periods in terms of the questions and answers about the Zika virus on Yahoo! Answers. The question and answer frequencies of the three categories versus the four periods are listed in Table 7. In the table, “Q” and “A” represent the number of questions and the number...
of answers, respectively. There were two peaks. They occurred in Period I (Q = 64, A = 476) when the disease broke out, and Period III (Q = 92, A = 501) when the Zika virus was under control and WHO declared that the Zika virus was no longer regarded as a “public health emergency” (November, 2016).

Over the four periods, the terms changed in the three categories. Some terms remained in all the categories while

![FIG. 2. The MDS display of Social impact of the Zika virus in Period I. [Color figure can be viewed at wileyonlinelibrary.com]](image1)

![FIG. 3. The MDS display of Consumer concerns about the Zika virus in Period I. [Color figure can be viewed at wileyonlinelibrary.com]](image2)
some terms only appeared in one category. The unique terms in each of the categories and overlapping terms in all the three categories are shown in Table 8. The numbers of the unique terms in the categories About the Zika virus, Social impact of the Zika virus, and Consumer concerns about the Zika virus were 30, 54, and 33 respectively. There were 38 overlapping terms in the three categories.

The terms defined the major common themes of the posts related to the Zika virus on Yahoo! Answers. The terms in a category changed over the observed periods. Some new terms were added to the category and some existing terms were removed from the category. The term number changes between two adjacent periods in the major categories over the four periods are illustrated in Table 9.

FIG. 4. The MDS display of About the Zika virus in Period II. [Color figure can be viewed at wileyonlinelibrary.com]
FIG. 5. The MDS display of Social impact of the Zika virus in Period II. [Color figure can be viewed at wileyonlinelibrary.com]

FIG. 6. The MDS display of Consumer concerns about the Zika virus in Period II. [Color figure can be viewed at wileyonlinelibrary.com]
For the comparison between Period I versus Period II, the newly added terms are mainly related to the Olympic Games, the impact of the Zika virus on holidays and travel, and health concerns. For the comparison between Period II and Period III, the new terms are primarily associated with related diseases, public reactions to the Zika virus, and the impact of the Zika virus on mothers and their new babies, and politics. For the comparison between Period III to Period IV, the new terms focus on Zika virus prevention.

Discussion and Implications

Comparisons With Previous Studies

The findings can be compared with other similar studies on social media from multiple perspectives.

Themes. The generated themes in studies on different health topics revealed different characteristics of the topics. Prior studies have sought to classify questions about different health-related topics posted on social Q&A websites. Bowler et al. (2012) concluded five overarching themes and 11 sub-themes for questions about eating disorders: Seeking Information (Factual, Diagnosis, Treatment or Intervention), Seeking Emotional Support (Validation, Seeking Comfort), Seeking Communication (Conversation Starters, Deep Talk),

TABLE 3. The MDS result of Period II.

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Causes and symptoms (11)</td>
<td>Zika, virus, bite, mosquito, head, sick, kill, infect, control, people, medicine south, America, unit, state, country, travel, plan, report, pregnant, contact</td>
<td></td>
</tr>
<tr>
<td>Transmission and risks (10)</td>
<td>disease, new, status, provide, poison, medicine, coast, water, pollute, health, Brazil, Olympic, athlete, game, country, cloth, cause, cancel, economy, threat, people, government, Christmas, holiday, vote, ISIS, fight, fly, liberty, war, German</td>
<td></td>
</tr>
<tr>
<td>Social events and travel (31)</td>
<td>Vaccine, Zika, virus, problem, south, America, mosquito, live, pregnant, foreign, born, baby, small, head, deform, sick, kill, infect, spread, attack, blasphemy</td>
<td></td>
</tr>
<tr>
<td>Protection and surveillance (21)</td>
<td>Prevention and solutions (15)</td>
<td>Christmas, holiday, avoid, cloth, bite, mosquito, develop, vaccine, pregnant, transmit, semen, scared, sick, problem, body</td>
</tr>
<tr>
<td>Attitudes and opinions (23)</td>
<td>God, Christ, abort, kill, live, affect, cause, control, government, health, head, born, baby, woman, people, sex, Brazil, south, America, country, Zika, virus, disease</td>
<td></td>
</tr>
</tbody>
</table>

For the comparison between Period I versus Period II, the newly added terms are mainly related to the Olympic Games, the impact of the Zika virus on holidays and travel, and health concerns. For the comparison between Period II and Period III, the new terms are primarily associated with related diseases, public reactions to the Zika virus, and the impact of the Zika virus on mothers and their new babies, and politics. For the comparison between Period III to Period IV, the new terms focus on Zika virus prevention.

FIG. 7. The MDS display of About the Zika virus in Period III. [Color figure can be viewed at wileyonlinelibrary.com]
FIG. 8. The MDS display of Social impact of the Zika virus in Period III. [Color figure can be viewed at wileyonlinelibrary.com]

FIG. 9. The MDS display of Consumer concerns about the Zika virus in Period III. [Color figure can be viewed at wileyonlinelibrary.com]
TABLE 4. The MDS result of Period III.

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory (number of terms)</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Causes and symptoms (21)</td>
<td>Zika, virus, infect, bite, cause, test, syndrome, sick, troll, head, defect, pregnant, woman, mother, born, baby, people, avoid, week, news, worry problem, disease, transmitt, outbreak, affect, spread, Africa, Nile, west, America, south, Brazil, Aedes, mosquito, fly, blood, kill, Ebola, Dengue, clean, silly</td>
</tr>
<tr>
<td></td>
<td>Transmission (21)</td>
<td>Brazil, Olympic, RIO, event, city, game, crime, attack, water, week new, born, baby, pay, support, head, microcephaly, live, sad, government, congress, foder, president, stuff, pregnant, woman, mother, abort, legal, idiot, medicine, control, conserve, infect, outbreak, protect, foreign, safe, policy, win, vote, hope, kill, mosquito, Trump, liberty, terror, Florid, Mexico, state</td>
</tr>
<tr>
<td>B</td>
<td>Social events and travel (10)</td>
<td>Brazil, Olympic, RIO, event, city, game, crime, attack, water, week new, born, baby, pay, support, head, microcephaly, live, sad, government, congress, foder, president, stuff, pregnant, woman, mother, abort, legal, idiot, medicine, control, conserve, infect, outbreak, protect, foreign, safe, policy, win, vote, hope, kill, mosquito, Trump, liberty, terror, Florid, Mexico, state</td>
</tr>
<tr>
<td></td>
<td>Activities (40)</td>
<td>Brazil, Olympic, RIO, event, city, game, crime, attack, water, week new, born, baby, pay, support, head, microcephaly, live, sad, government, congress, foder, president, stuff, pregnant, woman, mother, abort, legal, idiot, medicine, control, conserve, infect, outbreak, protect, foreign, safe, policy, win, vote, hope, kill, mosquito, Trump, liberty, terror, Florid, Mexico, state</td>
</tr>
<tr>
<td></td>
<td>Responses (28)</td>
<td>Zika, virus, disease, news, spread, Hillary, Obama, race, GOP, republic, parenthood, bill, plan, fund, blame, shame, hate, fault, south, America, country, area, travel, sick, fight, cause, problem, people</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cause, fault, God, Christ, religion, atheism, water, problem, small, head, sick, symptom, medicine, Zika, virus, disease, develop, live, born, baby, adult, people, worry, scared</td>
</tr>
<tr>
<td>C</td>
<td>Attitudes and opinions (24)</td>
<td>cause, fault, God, Christ, religion, atheism, water, problem, small, head, sick, symptom, medicine, Zika, virus, disease, develop, live, born, baby, adult, people, worry, scared</td>
</tr>
<tr>
<td></td>
<td>Prevention (12)</td>
<td>pregnant, woman, mother, week, Christmas, Florida, insect, avoid, repel, sex, drink, blood</td>
</tr>
<tr>
<td></td>
<td>Solutions (21)</td>
<td>pregnant, woman, mother, week, Christmas, Florida, insect, avoid, repel, sex, drink, blood</td>
</tr>
</tbody>
</table>

Seeking Self-Expression (Confession, Reflection), and Seeking Help to Complete a Task (Homework Help, Manuscript Ideas). Zhang and Zhao (2013) revealed 12 topics about diabetes: Cause & Pathophysiology, Sign & Symptom, Diagnosis & Test, Organ & Body Part, Complication & Related Disease, Medication, Treatment, Education & Info Resource, Affect, Social & Culture, Lifestyle, and Nutrient. The shared themes among these studies are diagnosis and treatment. Each health-related topic, however, has its unique focus. For example, people have asked confession questions for eating disorders, nutrient questions for diabetes, and transmission questions for the Zika virus. The unique subtheme discovered in this study was social events and travel. Because the Zika virus is transmissible, it does affect people’s everyday lives. Traveling to an area affected by the Zika virus can be dangerous. Therefore, it is not surprising that it would affect a large number of people.
FIG. 11. The MDS display of Social impact of the Zika virus in Period IV. [Color figure can be viewed at wileyonlinelibrary.com]

FIG. 12. The MDS display of Consumer concerns about the Zika virus in Period IV.
social event such as the Olympic Games in a country dealing with the virus and could cause public panic.

**Temporal analysis.** The Zika topic is time-sensitive, while the diabetes, eating disorder, and sexually transmitted diseases-related topics did not follow a temporal pattern (Bowler et al., 2013; Oh, 2015; Zhang & Zhao, 2013). The posts on the Zika virus topic primarily occurred during its outbreak period and decreased dramatically after the outbreak. This topic was time-sensitive and suggests that if consumers search for questions and answers for a time-sensitive topic like the Zika virus topic, time is an important factor.

The temporal analysis results showed that the number of questions and answers on the Zika virus varied in different time periods and had a close relationship with its impact on society. The outbreak of the Zika virus resulted in a surge of posts on Yahoo! Answers. The temporal analysis results illustrated that there were two peaks in terms of the posts on the Zika virus during the investigated period. These findings corroborated similar results revealed in the study of the Zika virus on other social media: Twitter, YouTube, and Wikipedia (Bragazzi et al., 2017). Here, two peaks of public responses to the Zika virus outbreak were also discovered. The first peak was directly triggered by the Zika virus outbreak, while the second peak was associated with the Olympic Games. It suggests that a transmissible disease like the Zika virus can result in public interest on social media. The degree to which the disease has an impact on a social event would be reflected in social media. Public health organizations can use social media to provide health consumers with necessary information such as disease prevention after a contagious disease breaks out or to warn social event organizers of the possible risks in affected areas.

**Contents.** Just as there were differences in interest between health topics, there were also some subtle differences in concerns between health consumers and health professionals. The CDC (2016a) listed the most common symptoms of the Zika virus as fever, rash, headache, joint pain, conjunctivitis (red eyes), and muscle pain. Al-Qahtani, Nazir, Al-Anazi, Rubino, and Al-Ahmad (2016) reported that the most common symptoms for people infected with the Zika virus were rash, fever, arthralgia, and conjunctivitis. The findings of this study revealed that symptoms for the Zika virus included sickness (that is, vomiting), fever, and microcephaly (small/deformed head). It is clear that microcephaly was heavily discussed on Yahoo! Answers, but it is not included on the CDC list. Similarly, Mishra and Behera (2016) summarized the transmission methods of the Zika virus through sexual transmission, blood transfusion, and perinatal transmission. In contrast, health consumers in this study referred to mosquito bites, sex or blood transmissions, and mother to child transmission. These users did not identify the transmission of the Zika virus through laboratory and healthcare setting exposure reported by the CDC (2016b).

There are also differences in the quality of information found on social media. Social media such as a public Q&A website is a platform where users can interact with each other on topics of interest. Information on social media is not reviewed by domain experts like a peer-reviewed journal. Therefore, it can include information of varying quality. It is not surprising that pseudoscientific claims on the Zika virus were found on some social media (Dredzea, Broniatowski, & Hilyard, 2016).

**Sentiment Analysis**

The outbreak of the Zika virus resulted in public panic around the world. Sentiment analysis revealed public emotional responses to the outbreak. The investigation of public
sentiment changes during the four periods adds another dimension to the study.

Lexalytics was used to analyze the sentiment expressed within the posts. An item was treated as an input data unit for the sentiment analysis. Each item contained one or more sentences that the users created. Lexalytics identified the emotive phrases within each source text and then generated a sentiment score ranging from $-2$ to $+2$ for a processed item.

TABLE 7. The question and answer frequencies of the three categories versus the four periods.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Period I</th>
<th></th>
<th>Period II</th>
<th></th>
<th>Period III</th>
<th></th>
<th>Period IV</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>About the Zika virus</td>
<td>Q</td>
<td>A</td>
<td>Q</td>
<td>A</td>
<td>Q</td>
<td>A</td>
<td>Q</td>
<td>A</td>
<td>405</td>
</tr>
<tr>
<td>Social impact of the Zika virus</td>
<td>27</td>
<td>204</td>
<td>7</td>
<td>42</td>
<td>29</td>
<td>130</td>
<td>9</td>
<td>29</td>
<td>72</td>
</tr>
<tr>
<td>Consumer concerns about the Zika virus</td>
<td>17</td>
<td>120</td>
<td>15</td>
<td>84</td>
<td>43</td>
<td>264</td>
<td>6</td>
<td>22</td>
<td>81</td>
</tr>
<tr>
<td>Subtotal</td>
<td>64</td>
<td>476</td>
<td>36</td>
<td>200</td>
<td>92</td>
<td>501</td>
<td>17</td>
<td>60</td>
<td>209</td>
</tr>
</tbody>
</table>

TABLE 8. The unique terms and overlapping terms in the major categories over the four periods.

<table>
<thead>
<tr>
<th>About the Zika virus</th>
<th>Social impact of the Zika virus</th>
<th>Consumer concerns about the Zika virus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique terms</td>
<td>Athlete, attack, bill, blasphemy, cancel, city, coast, congress, conserve, crime, dark, economy, event, feder, fight, Florid, foreign, form, fund, game, German, GOP, hate, Hillary, idiot, ISIS, legal, liberty, Obama, Olympic, parenthood, pay, poison, policy, pollut, pregnant, president, protect, race, republic, RIO, sad, safe, shame, solution, spray, status, stuff, support, terror, threat, Trump, wall, win.</td>
<td>Adam, atheism, bomb, Catholic, Christ, cure, curse, discover, DNA, drink, emerge, epidemic, evolution, Florida, God, idea, insect, Jesus, modify, mutate, perfect, plague, punish, religion, repel, research, Satan, season, semen, serious, suffer, world, zombie.</td>
</tr>
<tr>
<td>Overlapping terms</td>
<td>Aedes, Africa, Caribbean, clean, company, concern, condition, contact, Dengue, fetus, fever, media, merge, mild, Modern, newborn, Nile, product, report, scapegoat, shrink, silly, similar, syndrome, Tdap, temperature, troll, tropic, Uganda, west.</td>
<td>affect, America, baby, bite, blame, born, Brazil, cause, control, country, defect, disease, head, health, infect, kill, live, medicine, microcephaly, mosquito, mother, new, people, plan, pregnant, prevent, problem, sick, small, south, spread, state, vaccine, virus, week, woman, worry, Zika.</td>
</tr>
</tbody>
</table>
An analysis of variance (ANOVA) test was conducted to test if there are significant differences across the four periods in terms of the average sentiment. There were significant differences across the four periods in terms of the post sentiment ($F(3, 1334) = 3.999, p = .008 < .05$). To identify the pairs that caused the rejection, a Tukey HSD test was performed. The results indicated very strong evidence of sentiment differences between Periods I and II.

Figure 14 displays the boxplots of the sentiment scores during the four periods. The medians and spread of the data are illustrated in the figure. In Figure 14, each boxplot represents a period. The crosses correspond to the means.

The results in Figure 14 showed that the means of the sentiment scores ($-0.146, -0.058, -0.121,$ and $-0.108$) in the four periods were negative. The negative sentiment scores in conjunction with the revealed terms related to panic (for instance, danger, worried, concern, sad, scared, epidemic, kill) confirmed that the Zika virus outbreak did cause some level of public panic. The mean of the sentiment scores was lowest in Period I at the beginning of the Zika virus outbreak. It improved during Period II. Then it plummeted again in Period III, coinciding with the 2016 Olympic Summer Games. It is interesting that the two periods with the lowest sentiment scores coincided with the two periods of the post peaks.

A Pearson’s $r$ correlation analysis was also conducted to test the correlations between the questions and the associated answers in terms of the sentiment. There was a small to medium positive correlation between the questions and the associated answers in terms of the sentiment ($r = 0.141, N = 209, p = .041 < .05$). This suggests that the questions that showed more positive sentiment tended to receive more positive answers.

**Implications**

The results of the study help both health consumers and medical professionals better understand health consumers’ concerns and information needs about the Zika virus. The findings showed that health consumers on Yahoo! Answers were interested not only in the Zika virus symptoms, prevention, protection, and treatment but also in religion, social events like the Olympic Games, travel, and emotions.

The consumer-oriented schema can be used to organize the Zika virus-related information in Zika-related websites, digital libraries, information repositories, and databases. For instance, each item on Wikipedia has a table of contents, “see also” internal links, references, external links, and so on. The schema can be used to organize the table of contents of the corresponding Wikipedia site, making user navigation more effective and efficient.

The discovered terms relationships and patterns can be employed to enrich the existing Zika virus-related medical thesaurus entries, subject headings, and classification systems. These systems include MeSH (Medical Subject Headings) and MedlinePlus Health Topics. In the current MeSH version, the term “Zika virus” does not have any related

![FIG. 14. Boxplots of the sentiment scores in the four periods. [Color figure can be viewed at wileyonlinelibrary.com]](image_url)
the Zika virus outbreak in 2016.

A Zika virus schema and term connections were revealed based on expressed consumer health information needs. Two peaks in the Zika virus-related posts on Yahoo! Answers were identified in the temporal analysis. The first peak was directly triggered by the Zika virus outbreak, while the second peak was associated with the 2016 Olympic Summer Games. The Zika virus had a strong impact on social events like the Olympic Games and people’s travels, especially pregnant women.

The number of questions about the Zika virus fluctuated greatly during the different periods. We found there were significant differences between the revealed three categories and defined four periods in terms of the number of items about the Zika virus on Yahoo! Answers.

There are limitations to this study. First, although the data collection covered 1 year, and it included the entire run of the Zika virus from outbreak to its end, the sample size was relatively small. Second, only Yahoo! Answers was investigated. Consumers health needs on Zika virus on other health specific or related social media platforms or in different languages might include more consumer interaction behaviors and information needs.

The consumer-oriented subject schema can be applied to a variety of situations ranging from the Zika virus-based ontology to navigation guidance of a Zika virus-based portal. The findings of this study provided clusters of relevant terms on the Zika virus. These can be used to enrich existing health consumer-based vocabulary or thesaurus and enhance search engine effectiveness.

Conclusion

This study investigated health consumers’ concerns and information needs about the Zika virus during the period of the Zika virus outbreak in 2016.

A Zika virus schema and term connections were revealed based on expressed consumer health information needs. Two peaks in the Zika virus-related posts on Yahoo! Answers were identified in the temporal analysis. The first peak was directly triggered by the Zika virus outbreak, while the second peak was associated with the 2016 Olympic Summer Games. The Zika virus had a strong impact on social events like the Olympic Games and people’s travels, especially pregnant women.

The number of questions about the Zika virus fluctuated greatly during the different periods. We found there were significant differences between the revealed three categories and defined four periods in terms of the number of items about the Zika virus on Yahoo! Answers.

There are limitations to this study. First, although the data collection covered 1 year, and it included the entire run of the Zika virus from outbreak to its end, the sample size was relatively small. Second, only Yahoo! Answers was investigated. Consumers health needs on Zika virus on other health specific or related social media platforms or in different languages might include more consumer interaction behaviors and information needs.

The consumer-oriented subject schema can be applied to a variety of situations ranging from the Zika virus-based ontology to navigation guidance of a Zika virus-based portal. The findings of this study provided clusters of relevant terms on the Zika virus. These can be used to enrich existing health consumer-based vocabulary or thesaurus and enhance search engine effectiveness.

References


Li, L., He, D., Jeng, W., Goodwin, S., & Zhang, C. (2015). Answer quality characteristics and prediction on an academic Q&A Site: A case


