IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 33, No. 5, October, 2019, pp. 359-365
Journal homepage: http://iieta.org/journals/ria

Prediction of the Dissemination of Health News on Microblogging Sites Based on Ample Feature Selection and Support Vector Machine

Jiayin Pei, Peng Shan*

School of Business, Jiangnan University, Wuxi 214122, China

Corresponding Author Email: shanp@jiangnan.edu.cn

https://doi.org/10.18280/ria.330505

Received: 17 April 2019
Accepted: 21 July 2019

Keywords:
feature selection, binary classification, news dissemination, support vector machine (SVM)

ABSTRACT

As a social networking service, microblogging sites provide an open platform facilitating the sharing and discussion of valuable news. This paper identifies the influencing factors of the dissemination of health news posts on Weibo, the leading microblogging site in China. The effects of these factors were tested with 863 news posts, all about public health issues. The content features, author features and a social feature of each post were evaluated, and collected into a set of ample, diverse features that characterize widely disseminated posts. In addition, the support vector machine (SVM) was adopted to differentiate between widely disseminated posts and normal posts, and compared with several classification methods through an experiment on microblog posts of health news. The results show that the SVM greatly outperformed the contrastive methods in predicting the dissemination trends of such news. The research results inform crisis managers about the public reaction towards specific news on public health issues, shedding important new light on news dissemination.

1. INTRODUCTION

Microblogging sites provide users with a medium to broadcast small, digital contents. These sites are emerging as a fast communication channel for gathering and spreading breaking news [1]. Each day, microblog posts report all kinds of news, ranging from local issues to global events [2]. Monitoring and analyzing this rich content can yield unprecedentedly valuable information, blessing users and organizations with actionable knowledge [3].

For time-sensitive professions that constantly compete for attention (e.g. journalism, crisis management and news recommendation), it is extremely important to accurately estimate the spreading range of a news post on a microblogging site. However, the accurate estimation is by no means an easy task. For instance, the features related to news spreading are hard to identify. Quite a few studies have examined the factors related to risk information dissemination. Nonetheless, scholars engaged in text mining only consider words as features, which are inefficient and hard to interpret [4]. What is worse, the literature on features related to information dissemination contains many inconsistent findings [5-8]. Some scholars believed that long messages are more persuasive than short ones, while some disagreed with this belief. In fact, context specific features should be highlighted to avoid misrepresentation of the posts. Another difficulty lies in selecting the optimal prediction algorithm. Crisis events are urgent, uncertain and constantly changing. Once a crisis occurs, it takes considerable efforts to obtain efficient and reliable outcomes, such as producing more suitable features and utilizing more accurate prediction algorithms.

To solve the above difficulties, this paper explores the features important to microblog users in the search and dissemination of information. An ample subset of features was identified to differentiate between widely disseminated news posts from normal ones. Then, an optimal algorithm was selected to predict the dissemination trends. Since it is impossible to identify a particular feature that applies to all situations, the research focus is placed on public health issues, which have received much attention recently.

2. METHODOLOGY

2.1 Identification of influencing factors of news dissemination

Microblog posts that are widely disseminated often resonate with mass concern and may lead to public crisis. Therefore, it is of great interest to understand the causes of the proliferation of health news. The causal analysis is actually a binary classification problem to separate the normal posts from the widely disseminated posts. Before the separation, it is necessary to identify the influencing factors of news dissemination. The news should be represented correctly and completely by a set of rich and diverse features. Otherwise, the detected results will be imprecise or even meaningless. Previous research has shown that the content features of a microblog post, together with contextual information like author features and early-stage social features of the post, may influence the microblog users' decision on whether to forward the post. Therefore, the dissemination of a health news post may be affected by the combination of these three types of features [3].

(1) Content features

It is widely agreed that content features bear on information dissemination. However, there is no consensus over the conceptualization or operationalization [9]. One reasonable conceptualization is given by persuasion theory: attitude

valance and persuasive strength influence information dissemination in the form of persuasive information.

The traditional studies on persuasion have recognized attitude valance, positive or negative, as the key to persuading others. All kinds of decisions, from which candidate to support to which news message to share, are subjected to attitude. Persuasive strength is another common rule for decision-making. It is commonly measured by its relevance, timeliness, accuracy, and comprehensiveness [10].

Assuming all the microblog posts released by mainstream outlets are accurate, this paper measures persuasive strength by relevance, timeliness and comprehensiveness. Relevance refers to how relevant the news is related to public health issues; timeliness stands for how up-to-date the news post is; comprehensiveness means the coverage or depth of the post.

Unlike the traditional linguistic analysis, this paper extends the metrics of persuasion strength to nonlinguistic features in social media. Similar to Twitter, most microblogging sites set a limit on the length of each post. For example, a Weibo post should not exceed 140 characters. Within the restricted length, users are allowed to insert emoticons and URLs, mention others by adding an @ before their usernames, and even attach images to their posts. All these nonlinguistic features have a substantial influence on persuasive effect. Therefore, the stylistic features of microblog were also selected to evaluate the persuasion strength, in addition to the traditional metrics. All the selected content features are listed in Table 1.

Considering the nearly free text and hybrid types of elements in microblog posts, the quantitative content analysis was adopted to explain the linguistic and nonlinguistic content features. Based on preset categories, quantitative content analysis can count and quantify the content of each post. In this way, the unstructured microblog data are transformed into relational forms that can be reasonably analyzed.

Table 1. The content features of microblog posts

	Name of feature	Value range
	Attitude valance	
X 1	Number of unique positive words	(-inf-0], (0-1], (1-2], (2-3], (3-inf)
X 2	Number of neutral emotions	(-inf-0], (0-1], (1-inf)
X 3	Number of negative emotions	(-inf-0], (0-1], (1-3], (3-inf)
X 4	Number of unique negative words	(-inf-3], (3-5], (5-6], (6-8], (8-inf)
	Persuasive strength	
	Relevance	
X 5	Frequency of “food safety” terms	(-inf-1], (1-2], (2-3], (3-4], (4-inf)
X 6	Frequency of “medical” terms	(-inf-1], (1-2], (2-3], (3-4], (4-inf)
X 7	Frequency of “environment protection” terms	(-inf-1], (1-2], (2-3], (3-4], (4-inf)
	Timeliness	
X 8	The hour the post was published	(-inf-9], (9-12], (12-15], (15-19], (19-inf)
X 9	Whether the post was published in peak hours?	(0-8], (8-18], (18-24]
	Comprehensiveness	
X 10	The length of the news post	(-inf-138], (138-144], (144-149], (149-152], (152-inf)
X 11	Number of sentences	(-inf-10], (10-12], (12-13], (13-15], (15-inf)
X 12	Does the news post contain images?	(0,1)
X 13	Number of URLs	(-inf-0], (0-1], (1-2], (2-5], (5-inf)
X 14	Title length (the length of content within square brackets)	(-inf-13], (13-17], (17-20], (20-23], (23-inf)
X 15	The frequency of figures	(-inf-1], (1-3], (3-6], (6-10], (10-inf)
X 16	Number of words appearing once in the news post	(-inf-23], (23-27], (27-30], (30-33], (33-inf)
X 17	Number of words appearing twice in the news post	(-inf-1], (1-2], (2-3], (3-4], (4-inf)
X 18	Number of words appearing three times in the news post	(-inf-0], (0-1], (1-2], (2-3], (3-inf)
X 19	Number of question marks	(-inf-0], (0-1], (1-2], (2-3], (3-inf)
X 20	Number of exclamation marks	(-inf-0], (0-1], (1-2], (2-3], (3-inf)
X 21	Number of suspension points	(-inf-1], (1-2], (2-3], (3-4], (4-inf)
X 22	Number of mentions (@username) of other users	(0,1,2)
X 23	Topic length (the length of content within a pair of hashtags)	(-inf-1], (1-3], (3-4], (4-5], (5-inf)

(2) Author features

Most Internet users receive news from the media they usually follow. Thus, the features of the author exert a significant impact on the information evaluation by microblog users. As a result, this paper takes account of how author features affect news dissemination. Previous research has demonstrated that the activeness, experience and authority of the author are heuristic factors that represent three kinds of perceptions of the author [11]. The three factors are adopted in this paper to measure the features of the author. Of course, these metrics alone are too simple. For example, “whether an author is verified as an elite member” is not enough to reflect the authority of the author. Therefore, this paper adopts more metrics for the author features. The additional metrics are listed in Table 2. For instance, it is assumed that users are more likely to repost a news released by an author with a high influence score, which is a rating by the microblogging site. The author’s authority is also reflected by the length of self-description and the length of verification information.

(3) Social feature

The forward decisions of microblog users are subjected to social influence. Normative social influence is a type of social influence leading to conformity. In social psychology, it is defined as the influence of others that leads us to conform in order to be liked and accepted by others. Hence, normative social influence can be a very powerful, motivator of behavior, and may overshadow the impact of the authority of the author. Nevertheless, there is little report on the substantive importance of social influence [6]. For microblog users, whether to forward a microblog post depends on their perception of social norms concerning whether to perform the behavior. Therefore, this paper posits that the new posts can be discriminated by social based information. The social feature selected for this research is shown in Table 2.

Table 2. The author features and social feature of microblog posts

	Name of feature	Value range
	Author features	
	Author's activeness	
X24	Number of news posts published since the creation of the account	(-inf-30664], (30664-35698], (35698-45942], (45942-55256], (55256-inf)
X25	Number of users followed by the author	(-inf-213], (213-262], (262-396], (396-722], (722-inf)
X26	Number of followers of the author	(-inf-227], (227-258], (258-321], (321-516], (516-inf)
X27	Number of news posts favored by the author	(-inf-10], (10-133], (133-356], (356-528], (528-inf)
X28	Does the author allow personal messages?	(0,1)
	Author's experience	
X29	Number of days since the creation of the account	(-inf-795], (795-1235], (1235-1680], (1680-1703], (1703-inf)
X30	Gender of the author	(male, female)
X31	The year of the creation of the account	(2009,2010,2011,2012)
	Author's authoritativeness	
X32	Length of self-description	(-inf-13], (13-24], (24-42], (42-62], (62-inf)
X33	Length of verification information	(-inf-8], (8-9], (9-11], (11-15], (15-inf)
X34	Number of new followers attracted by the news post	(-inf-6063755], (6063755-6755320], (6755320-11794494], (11794494-18455034], (18455034-inf)
	Social information	
X35	Number of forwards of the news post within 120min	(-inf-72], (72-156], (156-310], (310-651], (651-inf)

2.2 Data collection

The research data were collected from Weibo, the leading microblogging site in China. The news posts from twelve mainstream news outlets covering public health issues were monitored continuously. To ensure the diversity of data sources, the twelve news outlets were carefully selected, including four newspapers, two magazines, five online news sites and one TV news agency. In total, 863 microblog posts were collected, all of which related to public health issues. Details on the dataset can be found in the previous research of Pei [6,12].

Over the years, many thresholds have been designed to evaluate widely disseminated posts. However, few have explained the selection criteria of these thresholds. Pei et al. [6] explored the dissemination of microblog information based on the Pareto Principle or the 80/20 Rule, revealing that 80 % of all the public attention are attracted by roughly 20 % of the posts (posts forwarded more than 766 times). In the light of this finding, the authors set the threshold of forwards to 766, and divided the 863 microblog posts into two classes. The 238 posts that were forwarded 766 times or more were defined as widely disseminated posts and categorized as Class 1, while the remaining 625 posts were defined as normal posts and categorized as Class 2. Then, the two classes were compared to discover the key features of widely disseminated posts, and forecast the types of news posts that will be widely diffused.

2.3 Measurement

(1) Attitude valance

The attitude valance of a microblog post is generally measured through sentiment analysis, which mainly detects the emotional sentiment reflected in the text for various purposes. An important basis of sentiment analysis is to identify positive and negative words. In this paper, the attitude valance of a microblog post is evaluated by metrics like the number of positive words, the number of neutral emotions, the number of negative emotions and the number of negative words.

(2) Persuasive strength

Relevance, a feature of persuasive strength, describes how relevant a news post is to public health issues. It can be measured by the number of public health words appearing in the text [6]. Hence, this feature is measured here by the frequencies of "food safety" terms, "medical" terms and "environmental protection" terms.

Timeliness, another determinant of persuasive strength, was evaluated by two metrics: the hour the post was published and whether the post was published in peak hours.

Comprehensiveness was assessed by various stylistic features of the post, such as "does the news post contain images". Specifically, the order and co-occurrence of words were used to judge the comprehensiveness. This is realized with the aid of n-gram, a contiguous sequence of n words from a given sequence of text. The n-gram technique, especially 2-gram words, can effectively characterize a message [13], and improve the prediction of the popularity of news posts. Comprehensiveness was also evaluated by microblog specific features like the topic length and number of mentions. Similar to twitter, Weibo users can indicate the topic of each post with a pair of hashtags, and mention others by placing an @ before their usernames.

Overall, 23 contents features, 11 author features and 1 social feature were selected for further analysis.

2.4 Feature reduction

The huge number of features may complicate the binary classification problem, suppressing the efficiency of the learning algorithms. Therefore, the predictive ability of each selected feature was investigated using the information gain method, aiming to find out the features that best reflect the attributes of news posts [14]. The information gain can be interpreted as follows [15].

Let Z be a discrete class variable with m alternative values. Then, entropy H(Z) can be defined as:

$$H(Z) = -\sum_{i=1}^{m} P(z_i) \log_2 P(z_i) \quad (1)$$

For a given X with k alternatives, the conditional entropy of Z can be expressed as:

$$H(Z|X) = \sum_{j=1}^{k} H(Z|x_j) P(x_j) \quad (2)$$

The information gain method aims to select features based

on the information contribution related to the class variable, i.e. the amount of additional information about Z provided by X:

$$IG(Z;X) = H(Z) - H(Z|X) \quad (3)$$

After obtaining the information gain of each feature, a minimum limit was set to filter out the features whose information gain is below the limit.

2.5 Classification

After feature reduction, each news post was transformed into a number of features that are represented as numerals. This lays the basis for the application of machine learning methods [16-17]. Support vector machine (SVM) is one of the most popular machine learning algorithms for classification and regression problems. The core idea of the SVM is to separate different classes with a hyperplane. This algorithm has been proved to have excellent effect on linear data classification in many applications [18]. If the original data are only nonlinearly separable, the SVM can be coupled with kernel functions to map the original data the feature space to a higher-dimensional space, such as to separate the data linearly. This subsection briefly describes the SVM algorithm adopted for our binary classification problem.

For linearly separable data, the decision function can be written as:

$$\mathrm{f(x)} = \mathrm{wTx} + \varepsilon = 0 \quad (4)$$

where, w is the weight vector; ε is the bias; x is the dataset. The hyperplane described by formula (4) divides one space into two parts: a positive part for samples in the positive class (+) and a negative part for samples in the negative class (-). Since the problem is to determine the values of w and ε, so that the hyperplane can be as far as possible from all the samples. More specifically, the SVM algorithm sets up hyperplanes, HP_1 and HP_2, as follows:

$$\begin{aligned} HP_1 &\rightarrow w^T x_i + b = +1 \; for \; y_i = +1 \\ HP_2 &\rightarrow w^T x_i + b = -1 \; for \; y_i = -1 \end{aligned} \quad (5)$$

where, $w^T x_i + \varepsilon \geq +1$ gives the hyperplane for the positive class; $w^T x_i + \varepsilon \leq -1$ gives the hyperplane for the negative samples. The two equations in formula (5) can be combined into:

$$y_i(w^T x_i + b) - 1 \geq 0 \quad \forall_i = 1, 2, \ldots, \mathrm{n} \quad (6)$$

The SVM margin represents the sum of d_1 and d_2 as:

$$margin = d_1 + d_2 = \frac{2}{||w||} \quad (7)$$

where, d_1 and d_2 are the distance of the samples from the first and second hyperplanes, respectively. In the SVM algorithm, the margin width needs to be maximized as:

$$\begin{aligned} &min \frac{1}{2}||w||^2 \\ s.t. \; &y_i(w^T x_i + b) - 1 \geq 0 \quad \forall_i = 1, 2, \ldots, \mathrm{n} \end{aligned} \quad (8)$$

Combining the objective function ($min \frac{1}{2}||w||^2$) and the constraint $y_i(w^T x_i + b - 1 \geq 0)$, the binary classification problem can be formalized into the following Lagrange formula:

$$minL_p = \frac{||w||^2}{2} - \sum_I a_i(y_i(w^T x_i + b) - 1) = \frac{||w||^2}{2} - \sum_I a_i(y_i(w^T x_i + b) + \sum_{i=1}^N a_i \quad (9)$$

where, a_i is the Lagrange multiplier for x_i; L_p is the primary problem. The values of w, ε, and a that minimize L_p in formula (9) were calculated by differentiating L_p with respect to w and ε and setting the derivatives to zero as:

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow \mathrm{w} = \sum_{i-1}^N \alpha_i y_i x_i \quad (10)$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{i=1}^N \alpha_i y_i = 0 \quad (11)$$

Substituting formulas (10) and (11) into formula (9), the binary classification problem can be transformed as:

$$\begin{aligned} maxL_D &= \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ s.t. \alpha_i &\geq 0, \sum_{i=1}^N \alpha_i y_i = 0 \; \forall_i = 1,2, \ldots, N \end{aligned} \quad (12)$$

where, L_D is the dual form of L_p. Then, the values of w, ε, and α can be determined by finding out a solution to formulas (10)~(12). In the SVM, most of the α_i are zeros, an evidence of the sparseness of the algorithm. The nonzero α values are the samples closest to the hyperplane, corresponding to support vectors (SV_s). Hence, the SV_s achieved the maximum width margin.

For nonlinear data, the key of the SVM is to use a nonlinear mapping function, i.e. the kernel function, to map samples into a higher-dimensional linear space. After the mapping, linear classification can be performed in the higher-dimensional space. After kernel transformation by the kernel function $K_{(x, y)}$, the decision function becomes:

$$f_x = \mathrm{w} \cdot \Phi_x + \varepsilon \quad (13)$$

2.6 Optimization

Then, several components of the SVM algorithm was optimized, including the penalty factor C, the kernel function width q and the insensitive band loss function g. The commonly used optimization methods include particle swarm optimization (PSO), and genetic algorithm (GA). After that, the classification effect of our model was evaluated by mean absolute percentage error (MAPE), mean square error (MSE) and Theil's coefficient of inequality.

2.7 Comparative experiment

Finally, the SVM classification performance of the research data were compared with traditional intelligent algorithms like Naive Bayes classifier, logistic regression, J48 tree classifier, and radial basis function (RBF) network. Many scholars have utilized one of these methods to predict information cascades, but few explained the reason to choose a specific method [19-20]. That is why the authors decided to compare the effectiveness of all these methods and verify if the SVM

outperform the contrastive methods in predicting the dissemination of health news.

3. RESULTS ANALYSIS

3.1 Ample features

To predict the dissemination trend of a news post, the features whose information gain is above 0.01 were selected and ranked in descending order of information gain (Table 3). As shown in Figure 3, the number of forwards of the news post within 120min, being a social feature, provides the highest information gain (0.6096). This means social features can greatly influence the dissemination of health news. This makes sense in the dissemination of risk information, in that a risk information reposted by numerous users early on can spread to more users in future. Therefore, it is meaningful to include social features of microblog news posts to existing models of information dissemination. Recommendation systems can also utilize social feature like the bandwagon effects [21].

Contrary to a previous study [22], the information gains indicate that author features are crucial to health news dissemination on microblogging sites. The posts published by an active, experienced and authoritative author can propagate widely across the network.

In addition, the content features directly bear on the dissemination trend of microblog news posts. This agrees with the previous research. For example, a title helps to convey the key message of the news, making the post more attractive to users. The proliferation of a public health news post is positively correlated with its comprehensiveness, i.e. the presence of a proper title and the massive use of emoticons, especially neutral emotions.

Table 3. List of the selected features

Number	Name of feature	Information gain
35	Number of forwards of the news post within 120min	0.6096
29	Number of days since the creation of the account	0.1913
32	Length of self-description	0.1408
31	The year of the creation of the account	0.1383
34	Number of new followers attracted by the news post	0.1271
26	Number of users followed by the author	0.0901
25	Number of followers of the author	0.0787
33	Length of verification information	0.0713
24	Number of news posts published since the creation of the account	0.0662
27	Number of news posts favored by the author	0.0448
14	Title length	0.0180
9	Whether the post was published in peak hours?	0.0176
2	Number of neutral emotions	0.0167
30	Gender of the author	0.0153
20	Number of exclamation marks	0.0120
23	Topic length	0.0103

3.2 Optimal classification algorithm

The final classification results of our SVM algorithm are listed in Table 3. For comparison, Naive Bayes classifier, logistic regression, J48 tree classifier, and RBF network were extracted from the Weka library, and also tested on the set of 16 features with high information gain. The classification performance of the SVM and these contrastive algorithms on microblog posts of health news are compared in Table 4. Each method was tested by 10-fold cross-validation. The performance was evaluated pairwise by precision, and F-score.

It can be seen from Table 4 that most of the contrastive methods achieved an average accuracy of no more than 80 %, while the SVM realized the highest Class 1 F-score (0.841), Class 1 precision (0.969), and Class 1 recall (0.769). Therefore, our algorithm outperformed all the contrastive methods in determining whether a news post will be widely disseminated. Although J48 tree classifier achieved relatively high Class 1 precision (0.967), it failed to obtain a sufficiently high Class 1 recall. The research results are consistent with the conclusion in previous studies that the SVM is better than standard machine learning algorithms.

Table 4. Prediction performance of different algorithms

Classification methods	Class 1 precision	Class 1 recall	Class 1 F-score	Weighted average precision
Naive Bayes	0.655	0.789	0.72	0.844
Logistic regression	0.879	0.765	0.818	0.905
J48 tree classifier	0.967	0.731	0.833	0.923
RBF network	0.654	0.748	0.698	0.831
SVM	0.969	0.769	0.841	0.921

4. CONCLUSIONS

Social media like microblogging sites have already developed into 24/7 dissemination platform of real-world events. Despite the research efforts in information dissemination, it remains a huge challenge to monitor or detect the risk information from social media services. For crisis managers, it is extremely important to understand the news propagation on microblogging sites and predict future crisis based on microblog news posts. Therefore, this paper explores deep into the features of microblog posts on health news and their dissemination trends.

This research makes several theoretical contributions. First, the authors identified the inconsistencies of conceptualization in the literature on content research, and re-conceptualized content features according to the persuasion theory. Second, this research incorporates the new realities of microblogging sites into the conceptualization of factors related to attitude strength. Third, this research highlights the importance of social information and the key role of author features in news evaluation.

There are also direct practical implications of our research. For example, the authors developed a set of ample, diverse features and an optimal algorithm to predict news dissemination. These results provide a valuable reference to researchers on information cascades. Furthermore, our approach targets specifically the news on public health issues. Thus, our analysis informs crisis managers about public reaction towards specific health news. Finally, our method is computationally feasible in near real-time scenarios and can be utilized to capture rapidly changing dynamics of microblog news dissemination.

ACKNOWLEDGEMENT

This paper is supported by National Natural Science Foundation of China (Grant No.: 71871106), Jiangsu Provincial Social Science Foundation (Grant No.: 18GLD014), Jiangsu Provincial University Philosophy and Social Science Foundation (Grant No.: 2018SJA0815), the Fundamental Research Funds for the Central Universities (Grant No.: 2019JDZD06) and Chinese University Research Foundation for Young Scholars (Grant No.: JUSRP11882).

REFERENCES

[1] Ucar, T., Culpan, M., Caskurlu, T., Karaman, M., Silay, M.S. (2018). The activity and discussion points of #Circumcision through Twitter; a microblogging platform. International Journal of Impotence Research, 30(5): 249-252. http://dx.doi.org/10.1038/s41443-018-0058-y

[2] Hasan, M., Orgun, M.A., Schwitter, R. (2019). Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. Information Processing & Management, 56(3): 1146-1165. http://dx.doi.org/10.1016/j.ipm.2018.03.001

[3] Akcura, T., Altinkemer, K., Chen, H.L. (2018). Noninfluentials and information dissemination in the microblogging community. Information Technology & Management, 19(2): 89-106. http://dx.doi.org/10.1007/s10799-017-0274-z

[4] Carrera, B., Jung, J.Y. (2018). SentiFlow: An information diffusion process discovery based on topic and sentiment from online social networks. Sustainability, 10(8): 2731. http://dx.doi.org/10.3390/su10082731

[5] Gursoy, D. (2019). A critical review of determinants of information search behavior and utilization of online reviews in decision making process (invited paper for 'luminaries' special issue of International Journal of Hospitality Management). International Journal of Hospitality Management, 76: 53-60. http://dx.doi.org/https://doi.org/10.1016/j.ijhm.2018.06003

[6] Pei, J.Y., Yu, G., Tian, X.Y., Donnelley, M.R. (2017). A new method for early detection of mass concern about public health issues. Journal of Risk Research, 20(4): 516-532. http://dx.doi.org/10.1080/13669877.2015.1100655

[7] Wu, L.R., Li, J.J., Qi, J.Y. (2019). Modeling information popularity dynamics based on branching process. Acta Physica Sinica, 68(7): 6. http://dx.doi.org/10.7498/aps.68.20181948

[8] Yan, J., Zhou, Y., Wang, S.Y., Li, J. (2019). To share or not to Share? Credibility and dissemination of electric vehicle-related information on WeChat: A moderated dual-process model. IEEE Access, 7: 46808-46821. http://dx.doi.org/10.1109/ACCESS.2019.2909072

[9] Aroean, L., Dousios, D., Michaelidou, N. (2019). Exploring interaction differences in Microblogging Word of Mouth between entrepreneurial and conventional service providers. Computers in Human Behavior, 95: 324-336. http://dx.doi.org/10.1016/j.chb.2018.10.020

[10] Cheung, C.M.K., Thadani, D.R. (2012). The impact of electronic word-of-mouth communication: A literature analysis and integrative model. Decision Support Systems, 54(1): 461-470. http://dx.doi.org/10.1016/j.dss.2012.06.008

[11] Zhang, L., Peng, T.Q., Zhang, Y.P., Wang, X.H., Zhu, J.J.H. (2014). Content or context: Which matters more in information processing on microblogging sites. Computers in Human Behavior, 31: 242-249. http://dx.doi.org/10.1016/j.chb.2013.10.031

[12] Pei, J., Yu, G., Shan, P. (2016). Social media coverage of public health issues in China: A content analysis of Weibo news posts. Information Technology: New Generations, 448: 111-120. https://doi.org/10.1007/978-3-319-32467-8_11

[13] Nerlich, B., Forsyth, R., Clarke, D. (2012). Climate in the news: How differences in media discourse between the US and UK reflect national priorities. Environmental Communication: A Journal of Nature and Culture, 6(1): 44-63. http://dx.doi.org/10.1080/17524032.2011.644633

[14] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. (2009). The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 11(1): 10-18. http://dx.doi.org/10.1145/1656274.1656278

[15] Shannon, C.E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3): 379-423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x

[16] Colombo, E., Forte, G., Rossignoli, R. (2019). Carry trade returns with support vector machines. International Review of Finance, 19(3): 483-504. http://dx.doi.org/10.1111/irfi.12186

[17] Kumar, B., Vyas, O.P., Vyas, R. (2019). A comprehensive review on the variants of support vector machines. Modern Physics Letters B, 33(25): 1950303. http://dx.doi.org/Artn195030310.1142/S0217984919503032

[18] Sowah, R.A., Kuuboore, M., Ofoli, A., Kwofie, S., Asiedu, L., Koumadi, K.M., Apeadu, K.O. (2019). Decision support system (DSS) for fraud detection in health insurance claims using genetic support vector machines (GSVMs). Journal of Engineering, 2019:

1432597. http://dx.doi.org/Artn 143259710.1155/2019/1432597
[19] Orkphol, K., Yang, W. (2019). Sentiment analysis on microblogging with K-means clustering and artificial bee colony. International Journal of Computational Intelligence and Applications, 18(3): 1950017. http://dx.doi.org/Artn195001710.1142/S146902681950 0172
[20] Vazquez, S., Munoz-Garcia, O., Campanella, I., Poch, M., Fisas, B., Bel, N., Andreu, G. (2014). A classification of user-generated content into consumer decision journey stages. Neural Netw, 58: 68-81. http://dx.doi.org/10.1016/j.neunet.2014.05.026
[21] Choi, S.M., Lee, H., Han, Y.S., Man, K.L., Chong, W.K. (2015). A recommendation model using the bandwagon effect for E-marketing purposes in IoT. International Journal of Distributed Sensor Networks, 2015: 1-7. http://dx.doi.org/10.1155/2015/475163
[22] Masip, P., Guallar, J., Suau, J., Ruiz-Caballero, C., Peralta, M. (2015). News and social networks: audience behavior. El Profesional de la Informacion, 24(4): 363. http://dx.doi.org/10.3145/epi.2015.jul.02