mathematics

MDPI

Article

A Novel Framework for Mining Social Media Data Based on Text Mining, Topic Modeling, Random Forest, and DANP Methods

Chi-Yo Huang [1,*], Chia-Lee Yang [2] and Yi-Hao Hsiao [2]

[1] Department of Industrial Education, National Taiwan Normal University, Taipei 106, Taiwan
[2] National Center for High-Performance Computing, Hsinchu 300, Taiwan; joy.yang@nchc.org.tw (C.-L.Y.); ihow@nchc.narl.org.tw (Y.-H.H.)
* Correspondence: cyhuang66@ntnu.edu.tw; Tel.: +886-2-7749-3357

Abstract: The huge volume of user-generated data on social media is the result of the aggregation of users' personal backgrounds, past experiences, and daily activities. This huge size of the generated data, the so-called "big data," has been studied and investigated intensively during the past few years. In spite of the impression one may get from the media, a great deal of data processing has not been uncovered by existing techniques of data engineering and processing. However, very few scholars have tried to do so, especially from the perspective of multiple-criteria decision-making (MCDM). These MCDM methods can derive influence relationships and weights associated with aspects and criteria, which can hardly be achieved by traditional data analytics and statistical approaches. Therefore, in this paper, we aim to propose an analytic framework to mine social networks, feed the meaningful information via MCDM methods based on a theoretical framework, derive causal relationships among the aspects of the theoretical framework, and finally compare the causal relationships with a social theory. Latent Dirichlet allocation (LDA) will be adopted to derive topic models based on the data retrieved from social media. By clustering the topics into aspects of the social theory, the probability associated with each aspect will be normalized and then transformed to a Likert-type 5-point scale. Afterwards, for every topic, the feature importance of all other topics will be derived using the random forest (RF) algorithm. The feature importance matrix will be transformed to the initial influence matrix of the decision-making trial and evaluation laboratory (DEMATEL). The influence relationships among the aspects and criteria and influence weights can then be derived by using the DEMATEL-based analytic network process (DANP). The influence weight versus each criterion can be derived by using DANP. To verify the feasibility of the proposed framework, Taiwanese users' attitudes toward air pollution will be analyzed based on the value–belief–norm (VBN) theory by using social media data retrieved from Dcard (dcard.tw). Based on the analytic results, the causal relationships are fully consistent with the VBN framework. Further, the mutual influences derived in this work that were seldom discussed by earlier works, i.e., the mutual influences between altruistic concerns and egoistic concerns, as well as those between altruistic concerns and biosphere concerns, are worth further investigation in future.

Keywords: social media mining; text mining; topic modeling; random forest (RF); decision making trial and evaluation laboratory (DEMATEL); multiple criteria decision making (MCDM)

Citation: Huang, C.-Y.; Yang, C.-L.; Hsiao, Y.-H. A Novel Framework for Mining Social Media Data Based on Text Mining, Topic Modeling, Random Forest, and DANP Methods. *Mathematics* **2021**, *9*, 2041. https://doi.org/10.3390/math9172041

Academic Editor: Radu Tudor Ionescu

Received: 31 May 2021
Accepted: 19 August 2021
Published: 25 August 2021

1. Introduction

Social media are web-based services that allow people, publics, and organizations to cooperate, link, network, and form communities. Such services allow users to easily generate, co-generate, adapt, share, and participate in web contents created by users [1]. In the past few years, social media have become a dominant part of daily life for most people, with enormous implications and impacts on regional, national, and global economies and political situations [1]. At the moment when the impacts of conventional media lessened, social media rapidly diffused into the world.

Social media breaks down the borders between the physical world and the virtual world. In the past several years, scholars have started to integrate social theories with algorithms to investigate how people (also referred to as social atoms) interact with each other and how communities (also referred to as social molecules) are formulated [2]. The exclusivity of the data retrieved from social media requires new data mining techniques; these social media mining techniques can effectively manipulate user-created content with rich social relationships [2]. Typical relationships include homophilic relationships (such as friendships on Facebook and following/follower relationships on Twitter) and relationships based on value homophily (such as retweets on Twitter, +1 on Google+, and "likes" on Facebook) [3]. These novel techniques are within the scope of social media mining, a rapidly evolving sub-domain of data mining. Generally speaking, social media mining refers to the analytic procedure of demonstrating, visualizing, analyzing, and deriving patterns from social media data [2].

Nowadays, social media have become the emphasis of numerous academic studies, basically because they touch the majority of people worldwide who can access mobile devices like cellular phones, tablets, and notebook computers [4]. Social media are a good source of data for big data analytics [5], so scholars or practitioners can have deeper understanding of user preferences, discover significant trends, analyze user behaviors, or investigate people's lifestyles [4]. In general, social media can provide the data required to analyze preferences, states, texts, images, etc. [4].

The exceptional accessibility of big data about human behaviors has significantly altered the world [6]. However, the data retrieved from social media sites are huge, related, noisy, extremely unstructured, and incomplete [7]. The scale and characteristics of the data retrieved from social media differ significantly from the data traditionally adopted by social scientists to develop theories [7]. Scholars also have to think about the feasibility of applying social theories on social media data [7]. Thus, investigators as well as practitioners are aggressively inventing and testing novel analytic techniques and decision-making methods to obtain insights into anthropological behavior and afford decision supports to handle important social problems [6].

The algorithmic revolution, which includes automatic data processing, machine learning, and natural language processing (NLP) techniques, has made it feasible to apply these big data. In spite of the impression one may get from the social media, much data processing has not been uncovered by existing techniques of data engineering and processing [8]. Therefore, investigations into the integration of social media, NLP, and other methods of data analytics will be very important for deriving novel implications of data retrieved from social media in general, and those data related to some specific theoretical framework in particular. Some scholars (e.g., Yang et al. [9]) have already adopted NLP with structural equation models and given insights into data retrieved from social media. Though the partial least squares structural equation modeling (PLS-SEM) based approach indeed derives meaningful results, the influence relationships among aspects and criteria can further be derived to give more meaningful insights.

Several multiple criteria decision making (MCDM) methods have been developed in the past few decades. These include the analytic hierarchy process (AHP) [10], the analytic network process (ANP), decision-making trial and evaluation laboratory (DEMATEL) [10], and the DEMATEL-based analytic network process (DANP) [11,12]. The AHP and the ANP have been used to measure the weights of the components of the structure by pairwise comparisons, and then to rank the alternatives in the decision. AHP structures a decision problem into a hierarchy with a goal, decision criteria, and alternatives, while the ANP structures it as a network. DEMATEL is a comprehensive method for building and analyzing a structural model involving causal relationships among complex factors. These methods have been applied widely to numerous decision-making problems, which include economics, management, engineering, environmental science, etc. These methods were adopted to derive the weights associated with certain aspects or criteria. Meanwhile, the influence relationships, as well as the influence weights, have further been proposed and

widely adopted. These MCDM-based methods can actually give insights into decision-making problems, e.g., the influence relationships and influence weights, which statistical methods-based analytic frameworks cannot afford. The integration of MCDM methods with big data analytics in general, and social media mining in particular, has been rare. However, their integration can indeed derive very different results compared to those methods that integrate big data analytics with a statistical analysis method, e.g., social media mining with PLS-SEM.

Data retrieved from social media usually contain meaningful information. However, few scholars have tried to analyze these data based on decision-making methods. A document usually contains numerous topics; according to Chen et al. [13], even a short document may contain multiple topics. These topics can serve as the criteria for a decision-making problem, and the problem is, by nature, a MCDM one. The influence relationships among the major variables in the social media data and the weights associated with these variables can be derived in order to provide meaningful insights. However, based on the authors' limited knowledge, very few scholars have tried to mine social media using MCDM methods. Although MCDM methods can potentially provide specific insights into the data retrieved from big data in general, and social media data in particular, few scholars have tried to propose analytic frameworks to address this research gap. Furthermore, almost no scholars have tried to propose an integrated framework to derive the influence relationships among the aspects of a theoretical framework. Thus, it is necessary to integrate information retrieved from social media sites into an established theoretical framework.

Therefore, in this paper, we aim to propose an analytical framework to mine a social network, analyze the meaningful information using decision-making methods based on a specific theoretical framework (e.g., the technology acceptance model or the value–belief–norm theory [14]), derive causal relationships among the aspects of the theoretical framework, and, finally, compare the causal relationships with a social theory.

First, social media sites will be trawled. The user-generated contents related to some specific social issue(s) will be retrieved. Then, the Latent Dirichlet allocation (LDA) technique will be adopted to derive topic models based on those data retrieved from social media. According to the probability associated with each topic, the topics will be clustered. Then, these topics will be classified into a specific aspect of a model of a social theory. To feed the probability of data into the computation, the probability associated with each aspect of the model of the social theory will be normalized using a Likert-type 5-point scale. Afterwards, for every topic, the random forest (RF) algorithm will be adopted to derive the feature importance of all other topics. The feature importance matrix will be transformed into the initial influence matrix of DEMATEL. The influence relationships can be derived, along with the influence weight versus each criterion, by using DANP. The consistency between the influence relation map (IRM) and the social theory model will be checked. Discrepancies will be derived, which can provide further insights regarding social phenomena. The contents generated by Taiwanese users regarding attitudes toward the air pollution problem will be retrieved from Dcard (www.dcard.tw, access on 1 July 2021) to verify the feasibility of applying social media data to the value–belief–norm theory proposed by Stern et al. [14]. For readers' convenience, a list of abbreviations and symbols introduced in this work are listed in Tables A1 and A2 in Appendix A.

The remainder of this paper is organized as follows: Section 2 reviews the relevant literature regarding the emergence of social media, the mining of social media, data-driven decision-making (DDD), past works on the integration of data analytics and MCDM methods, and research gaps. Research methods, which include the analytic process, topic modeling, RF, DEMATEL, and DANP, will be reviewed in Section 3. Section 4 presents the analytic results of text mining, topic modeling, cluster analysis, DEMATEL, and DANP. Finally, the results are discussed in Section 5. Section 6 concludes the whole work.

2. Literature Review

According to Kaplan and Haenlein [15], social media are the set of internet-based applications which are built upon the concepts and technology of Web 2.0; social media enable the generation and exchange of content generated by users [2]. Numerous classes of social media sites have been created. Typical examples include Facebook (for social networking), Twitter (for microblogging), YouTube (for video sharing), etc. [2]. Social media mining is an emerging interdisciplinary research field whose arena includes techniques from computer science, statistics, sociology, and ethnography [2]. DDD is a practice of decision-making, where decisions are based on data analytics instead of on intuitions only [8]. Better data provide more chances for enhanced decision-making results [16]. During the past few decades, MCDM methods have been developed and adopted for numerous applications. However, in the age of big data analytics, DDD based on MCDM methods has seldom been adopted in manipulating big data in general and social media data in particular. Thus, in this section, past works on the emergence of social media, social media mining, DDD, MCDM-based DDD, and research gaps will be reviewed. The literature will serve as the basis for developing the integrated framework consisting of social media mining and MCDM methods.

Social media is not based on a single technology. Instead, social media integrate wide-ranging techniques, which include numerous online services that augment the capability of mutual communication in the social environment that forms the organization [17]. The kernel of social media is grounded on the provision of high visibility and open participation [17]. For practical applications, social media provide features which allow seamless sharing, commenting, responding, syndicating and interacting with content (text, voice and video) and connecting with others, and following and interacting with their activity streams [15,18]. Thus, social media offer a flexible platform which is fundamentally organic, free-flowing, and constructed to enable dynamic and emergent feedback loops of communication within a social group [17].

Nowadays, social media platforms are typically applied in expressing opinions or viewpoints regarding social events, news, etc., everywhere, without any limitation of time. Future prediction is the great wish of mankind [19]. In order to meet this forecasting demand, many studies have correctly proven the importance of social media data (e.g., [10,20–22]). Therefore, during the past several years, scholars (e.g., [23,24]) have demonstrated numerous applications in the related fields of social science [19].

Social media mining refers to the process of characterizing, analyzing, and deriving important patterns from data retrieved from social media, which are the result of social interaction [2]. Social media mining is a multidisciplinary domain which includes techniques from computer science, data engineering, social science, and mathematics [5]. The exploration of social media by the above-mentioned techniques helps us understand the mutual interactions of users [2]. Further, interesting patterns, information diffusion, influence relationships, effective and efficient recommendations, as well as novel social behavior can be explored on social media sites [2]. DDD refers to data analytics-based decisions [8]. Good sources of data imply better opportunities for good decisions [16]. Novel digital techniques have greatly enhanced the quality and quantity of data available for decision-makers [16].

The advantages of DDD have been verified convincingly [8]. Brynjolfsson et al. have demonstrated how companies' performance can be enhanced by using DDD [8]. DDD is also related to better financial results [8]. DDD has been broadly applied in numerous domains such as medical science, environmental engineering, education, energy management, policy definitions, etc. [20].

Nowadays, people are facing complicated decision-making problems that are filled with tremendous information, which can describe diverse aspects of problems via different methods. For decision-makers, uncovering an idea solution to a decision-making problem is not easy [20]. A rational method to tackle this kind of problem is to analyze various aspects and then integrate the analyses to create final solutions to the problems [20]. This

choice is called MCDM [20]. During the past few decades, numerous works based on MCDM have been conducted to assist people in solving complicated problems [20].

Traditional MCDM methods such as the AHP, the ANP, DEMATEL, and the DANP have been widely adopted for many decision-making problems. The AHP proposed by Saaty [10] aims to derive the weights relating to each aspect and criterion of a decision-making method by assuming independence among these aspects and criteria. Saaty also proposed the ANP [21], which can derive the weights being associated with the aspects and criteria of a decision-making problem by releasing the assumptions of independence. DEMATEL, proposed by Gabus and Fontela [22] of the Battelle Geneva Institute, has been widely adopted to construct the influence relationships among the aspects and criteria of a MCDM problem. The DANP, a fusion of DEMATEL and the ANP, can easily derive the influence weights of each aspect and criterion of a MCDM problem based on the results of DEMATEL. The DANP simplifies the analytic procedure of the ANP-based methods and considers every influence relationship, while deriving the influence weights. In ANP-based methods, a threshold value is usually defined to avoid too much complexity in the structure of decision-making problems to be solved. From a traditional perspective, it is very reasonable to adopt these methods. However, in the era of big data, decision makers can further consider the possibility of incorporating big data into the decision-making process instead of relying on a very limited number of experts. In the age of big data analytics, data fill the whole analytic process of MCDM [20]. Therefore, generating reasonable solutions based on contemporary observations and past data has turned out to be a dominant and fascinating matter [20]. To resolve this problem, Fu et al. [20] proposed a DDD framework based on the MCDM method, which has become the focus.

Few scholars have tried to integrate machine learning algorithms and MCDM methods to tackle big data in general and social media data in particular. Recently, Yang et al. [23] used text mining methods to retrieve papers adopting deep learning—a subset of machine learning—algorithms, and MCDM methods in using big data. Limited results were retrieved from major academic databases, including ScienceDirect, ACM, IEEE, Springer, Taylor & Francis, and Wiley Online Library. Some of these works use the AHP to assess risks [24], such preparing a flood hazard susceptibility map [25]. However, as mentioned in the prior paragraph, the assumptions of independence among the aspects and criteria bias the results. Yasmin et al. [26] used intuitionistic fuzzy DEMATEL (IF-DEMATEL) and the ANP to analyze the capabilities of big data analytics for firms. However, they are not really dealing with big data. Meanwhile, the framework faces problems similar to those mentioned in the prior paragraph—the complicated survey procedure and the loss of valuable information due to the threshold definition.

Muruganantham and Gandhi [27] provide one of the few studies to incorporate social media data into a MCDM method. In their study, the Technique for Order Performance by Similarity to Ideal Solution (TOPSIS) was introduced to rank influencers in a given social media data set. However, no influence relationships, weights, or confirmation with theoretical frameworks could be provided due to the natural limitation of the TOPSIS, which aims to rank the alternatives only.

In general, in spite of the impression one may get from the media, much data processing that has not been uncovered by existing techniques of data engineering and processing. Therefore, investigations on the integration of social media, NLP, and other methods of data analytics will be very important for deriving novel implications of the data retrieved from social media in general, and the data related to a specific theoretical framework in particular. However, very few scholars have tried to do so, especially from the perspective of MCDM, which can derive influence relationships, which can hardly be achieved by traditional data analytics and statistical approaches. Therefore, in this paper, we aim to propose an analytic framework to mine social network, feed the meaningful information to MCDM methods based on a theoretical framework, derive causal relationships amongst the aspects of the theoretical framework, and finally compare the causal relationships with a social theory.

3. Research Methods

First, social media sites will be trawled. The user-generated contents related to some specific social issue(s) will be retrieved. After that, the LDA technique will be adopted to derive topic models based on those data retrieved from social media. According to the probability associated with each topic, the topics will be clustered. Then, these topics will be classified into a specific aspect of a social theory model. To feed the probability of data into the computation, the probability associated with each aspect of the model of the social theory will be normalized using a Likert-type 5-point scale. Next, for every topic, RF will be adopted to derive the feature importance of all other topics. The feature importance matrix will be transformed into the initial influence matrix of DEMATEL. The influence relationships can thence be derived. The influence weight versus each criterion can be derived by using DANP. The consistency between the IRM and the social theory model will be checked. Discrepancies will be derived, which can provide further insights regarding social phenomena. Below, the methods will be introduced. The three data analytic techniques, namely, topic modeling, hierarchical cluster analysis, RF, and DANP methods, will be introduced in the following subsections. These methods will be used to derive data from social media sites, derive latent topics, cluster these topics into theoretical frameworks, derive feature importances, and then feed these feature importances into DANP to derive meaningful implications. The proposed process consists of the following five steps (see Figure 1 below):

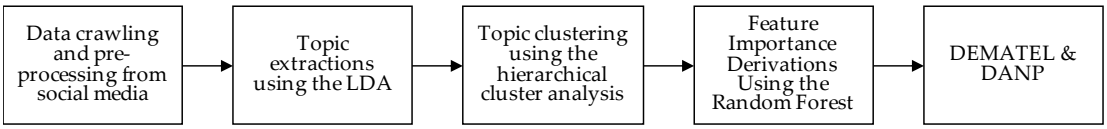

Figure 1. Research Framework.

3.1. Text Mining, Topic Model and LDA

Text mining was first proposed by Fledman et al. [28]. The term refers to the procedure of retrieving high-quality information from text, which includes structured, semi-structured, and unstructured text resources such as documents, videos, and images [29]. Text mining involves the extraction of information from text and the retrieving of text to derive rules and patterns [30]. Text mining also provides methods for analyzing and contextualizing massive volumes of information [31]. This, fundamentally, involves a quantitative method for analyzing (usually) big textual data; the techniques help accelerate knowledge discovery by drastically enhancing the amount of data to be analyzed [32].

One of the most popular methods of text mining is topic modeling. The method can effectively and systematically analyze many documents in a very short period of time. Among the topic modeling techniques, LDA [33], which is grounded on statistical distributions, is the most widely adopted. The basic assumption of LDA is an exchange among words and documents in a corpus, a bag of words. LDA recognizes semantically correlated words that appear at the same time in numerous documents in a corpus. After that, the topics of the words are inferred by humans as meaningful subjects. For example, the LDA assigns "gene," "DNA," "genetic," and "genetic" to topics that are interpreted as "genetic" [34].

Following, we define the terms and formulate the probabilistic model of a corpus based on the original definitions by Blei et al. [33]. A corpus D is defined as a collection of M documents. The number of words belonging to any one document d in the corpus is N_d, where $d \in \{1, \cdots, M\}$. The LDA algorithm models the corpus according to the below generative process based on the original definitions by Blei et al. [33] and Jelodar et al. [35]:

(a) Select a multinomial distribution φ_{t_p} for the topic t_p $(t_p \in \{1, \cdots, T\})$ from a Dirichlet distribution with parameter β.

(b) Select a multinomial distribution θ_d for document d $(d \in \{1, \cdots, M\})$ from a Dirichlet distribution with parameter α.

(c) For a word w_η ($\{\eta \in \{1, \cdots, N_d\}\}$) in document d,,

(i) Choose a topic z_η from θ_d.

(ii) Choose a word w_η from φ_{z_η},

where α is the per-document topic distributions; β is the per-topic word distribution; θ_d is the topic distribution for document d. θ_d is the topic distribution for the document d. The Dirichlet-multinomial pair for the corpus-level topic distributions is (α, θ), while the Dirichlet-multinomial pair for topic-word distributions is (β, φ).

In the above mentioned generative process, the words in the documents are observed variables while the others are latent variables (φ and θ_d) and hyper parameters (α and β). The probability of observed data (D) is computed and obtained as follows in Equation (1):

$$prob(D|\alpha, \beta) = \prod_{d=1}^{M} \int prob(\theta_d|\alpha) \left(\prod_{\eta=1}^{N_d} \sum_{z_{d_\eta}} prob\left(z_{d_\eta} \middle| \theta_d\right) prob\left(w_{d_\eta} \middle| z_{d_\eta}, \beta\right) \right) d\theta_d, \quad (1)$$

where z_{d_η} is the topic for the η-th word in document d and w_{d_η} is the specific word. Based on the above definitions, the probability of observed data will be derived using the LatentDirichletAllocation in the sci-kit learn Python toolkit [36].

3.2. The RF Technique

The RF method was proposed by Breiman [37] in 2001. It has been particularly effective as a classification and regression method. RF-based methods integrate some randomized decision trees and calculate the averages of predictions of these decision trees. These methods have demonstrated outstanding performance when the number of variables is much more than the number of observations [38]. Furthermore, the RF can be applied to large-scale problems, and can easily be modified to classify numerous arbitrary learning tasks by returning variable importance [38].

Based on the work of [39], the variable importance of a RF can be defined as follows. Assume a set $V = \{x_1, \cdots, x_p\}$ of categorical input variables and a categorical output y. Given a training sample S of n joint observations of $x_1, \cdots, x_p, y$ drawn from $P = \{x_1, \cdots, x_p, y\}$, let us define for any internal node t of a decision tree built from S:

The number of training samples in t as n_t;
The ratio of training samples in t as $p_r(t) = n_t/n$;
The impurity of node t as $i_p(t) = H(y|t)$;
The impurity reduction at node t as $\Delta i_p(t) = i_p(t) - (n_{tL}/n)\, i_p(t_L) - (n_{tR}/n) i_p(t_R)$,

where subscripts L and R are the left node and right node of the node t. In an ensemble of decision trees, the MDI importance of an input variable x_m is the sum of the weighted impurity reductions $p_r(t)\Delta i(t)$, for all nodes t where x_m is used, calculated as the averaged of all n_t trees in the ensemble:

$$\text{Imp}\ (x_m) = \frac{1}{n_T} \sum_{T_S} \sum_{t \in T_S : v(s_t) = x_m} p_r(t) \Delta i_p(s_t, t) \quad (2)$$

where T_S is a tree structure representing an input-output model and $v(t)$ is adopted to split node t [39].

A completely established, fully randomized decision tree is one in which every single node t is divided by means of a variable $x_{i_{RF}}$ selected uniformly at random (from among those nodes which have not been used at the parent nodes) into $\left|\aleph_{i_{RF}}\right|$ sub-trees (i.e., one for every possible value of $\aleph_{i_{RF}}$); the recursive construction ends when each one of the p variables has been used along the present branch [39].

The MDI importance of $x_m \in V$ for y as computed with an infinite ensemble of fully developed totally randomized trees and an infinitely large training sample is:

$$\text{Imp}\ (x_m) = \sum_{k_r=0}^{p-1} \frac{1}{C_p^{k_r}} \frac{1}{p-k_r} \sum_{B \in \mathcal{P}_k(V^{-m})}^{1} I(x_m; y|B), \tag{3}$$

where V^{-m} denotes the subset $V \backslash \{x_m\}$, $\mathcal{P}_{k_r}(V^{-m})$ is the set of subsets of V^{-m} of cardinality k_r, and $I(x_m; y|B)$ is the conditional mutual information of x_m and y given the variables in B [39].

For any ensemble of fully developed trees in asymptotic learning sample size conditions we have

$$\sum_{m=1}^{p} \text{Imp}(x_m) = I(x_1, \cdots, x_p, y) \tag{4}$$

$x_i \in V$ is irrelevant to y regarding V if and only if its infinite sample size importance, as computed with an infinite ensemble of fully developed totally randomized trees built on V for y, is 0 [39].

Let $V_R \in V$ be the subset of all variables in V that are relevant to y with respect to V. The infinite sample size importance of any variable $x_m \in V_R$ as computed with an infinite ensemble of fully developed totally randomized trees built on V_R for y is the same as its importance computed in the same conditions by using all variables in V [39].

Based on the above definitions, for every topic being derived in Section 3.1, the feature importance of all other topics will be derived using the RF algorithm with the RandomForestRegressor in the sci-kit learn Python toolkit [36]. The feature importance matrix will be transformed into the initial influence matrix of the DEMATEL, which will be introduced in the following Section 3.3.

The feature importance matrix $\boldsymbol{M}_F$ is defined as follows. In each column, the criteria importance will serve as the influence degree from a topic to some other specific topic. Further, each column of the transposed matrix will be normalized by the maximum element of the column. Then, every element will be multiplied by 5 for consistency with the Liker's 5-point scale adopted in later methods.

$$\boldsymbol{M}_F = \begin{bmatrix} 0 & I_{mp}(x_{2,1}) & \cdots & I_{mp}\ (x_{j_f,1}) & \cdots & I_{mp}\ (x_{p,1}) \\ I_{mp}\ (x_{1,2}) & 0 & \cdots & I_{mp}\ (x_{j_f,2}) & \cdots & I_{mp}\ (x_{p,2}) \\ \vdots & & \ddots & \vdots & & \vdots \\ I_{mp}\ (x_{1,i_f}) & I_{mp}\ (x_{2,i_f}) & \cdots & 0 & \cdots & I_{mp}\ (x_{p,i_f}) \\ \vdots & & & \vdots & \ddots & \vdots \\ I_{mp}\ (x_{1,p}) & I_{mp}\ (x_{2,p}) & & I_{mp}\ (x_{j_f,p}) & & 0 \end{bmatrix} \tag{5}$$

For each column, the largest element of the column, namely l_{j_f}, will be used to normalize the elements belonging to that column. Then, to be consistent with the Liker's 5-point scale, the normalized result will be multiplied by 5 as $\omega_{i_\omega j_\omega}$ of the $\boldsymbol{\Omega}$ matrix below. That is, $\omega_{i_\omega j_\omega} = I_{mp}\ (x_{p,i_\omega})/l_{j_\omega}$, where $l_{j_\omega} = 5 \cdot (\max_p I_{mp}\ (x_{j_\omega,h}))$, $h \in \{1, \cdots, p\}$.

$$\boldsymbol{\Omega} = \begin{bmatrix} 0 & \omega_{12} & \cdots & \omega_{1j_\omega} & \cdots & \omega_{1p} \\ \omega_{21} & 0 & \cdots & \omega_{2j_\omega} & \cdots & \omega_{2p} \\ \vdots & & \ddots & \vdots & & \vdots \\ \omega_{i_\omega 1} & \omega_{i_\omega 2} & \cdots & \omega_{i_\omega j_\omega} & \cdots & \omega_{i_\omega p} \\ \vdots & & & \vdots & \ddots & \vdots \\ \omega_{p1} & \omega_{p2} & & \omega_{pj_\omega} & & 0 \end{bmatrix}. \tag{6}$$

3.3. DEMATEL

DEMATEL was originally proposed by Gabus and Fontela [22] to solve complex world problems. It is based on the graph theory of discrete mathematics, and it can be used to derive the influence relationships among the criteria of a decision-making problem. Over the past years, DEMATEL has been widely adopted to solve numerous problems of policy definition, management (e.g., [40–42]), education (e.g., [43–46]) engineering (e.g., [47]), medical devices (e.g., [48]), and other social problems (e.g., [49]).

The basic DEMATEL formulas, by Tzeng and Huang [50], Yang et al. [40], and Huang et al. [47] are explained in the following procedure. First, the initial direct relation matrix (IDRM) can be formulated. Based on the $\boldsymbol{\Omega}$ matrix being derived by the RF, the influence of topic i_d on topic j_d, denoted as $a_{i_d j_d}$ in the IDRM, will be equal to $\omega_{i_d j_d}$ in the i_dth row and the j_dth column. Thus, $\boldsymbol{A} = \boldsymbol{\Omega}$, where $\boldsymbol{A} = [a_{i_d j_d}]$, $i_d, j_d \in [1, \cdots, T]$. Here, the row and column numbers equal to the number of topics T. Then, the IDRM will be normalized by multiplying the IDRM with a factor ρ using the Equation (7) below, i.e., $\boldsymbol{N}_R = \rho \boldsymbol{A}$, where the maximum row sum and the maximum column sum can be selected and ρ is equal to the smaller of the reciprocal of both numbers. That is,

$$\rho = \min\left\{1/\max_{i_d}\sum_{j_d=1}^{T} a_{i_d j_d}, 1/\max_{j_d}\sum_{i=1}^{T} a_{i_d j_d}\right\}, i_d, j_d \in \{1, 2, \ldots, T\}. \tag{7}$$

Then, the total relation matrix (TRM), $\boldsymbol{T}_R = \left[t_{i_d j_d}\right]_{T\times T}$, can be derived as: $\boldsymbol{T}_R = \boldsymbol{N}_R + \cdots + \boldsymbol{N}_R^{\varsigma} = \boldsymbol{N}_R(\boldsymbol{I}_d - \boldsymbol{N}_R)^{-1}$, where $\varsigma \to \infty$, $\boldsymbol{I}_d$ is the identity matrix. Then, the row sum and column sum vectors of the TRM can be derived as $\boldsymbol{r}$ and $\boldsymbol{c}$, respectively. The causal diagram or the IRM of all the aspects and topics can be derived by demonstrating the influence relationships, where $r_{i_d} + c_{i_d}$ and $r_{i_d} - c_{i_d}$ represent the horizontal and vertical axis of the topic.

3.4. The DANP

The DANP is an analytic method that integrates DEMATEL and the ANP proposed by Prof. Gwo-Hshiung Tzeng [11,12]. Traditionally, the ANP requires a pre-defined structure of the decision-making problem. Thus, decision makers may introduce the structure based on the IRM being derived by DEMATEL (refer to [41] for a typical example) or by other analytic methods. However, such work usually requires two or more iterations of collecting questionnaires, which wastes time and can be complicated. Respondents to the first iteration questionnaire may refuse to provide opinions for the second iteration questionnaire, which usually causes problems of inconsistency. Moreover, due to the complicated IRM derived by DEMATEL, a threshold value is usually required to screen the most important influence relationships inside the TRM. However, such screening usually filters out a lot of connections in the TRM. To overcome such limitations, the DANP feeds the IRM by DEMATEL into the ANP. By leveraging the super-matrix being proposed by Saaty in the ANP [21], the influence weights can be derived based on following procedures.

Based on the TRM ($\boldsymbol{T}_R$) derived in Section 3.3, the influence weights versus each topic can be derived by using the DANP method according to [42]. Let $\boldsymbol{T}_C$ be equal to the transposed matrix of the TRM, i.e., $\boldsymbol{T}_C = \boldsymbol{T}_R^t$. The TRM can be divided into m_s submatrices according to the topics belonging to the aspects. That is, $\boldsymbol{T}_C = \left[\boldsymbol{T}_{C_{i_S j_S}}\right]_{m_S \times m_S}$. The submatrices can be denoted as $\boldsymbol{T}_{C_{i_S j_S}} = \left[t_{i_u j_v}\right]_{i_{n_i} i_{n_j}}$, where $1 \le i_u \le i_{n_i}$ and $1 \le j_v \le i_{n_j}$. Here, n_i and n_j are the numbers of topics which belong to the i_Sth aspect, $\boldsymbol{D}_{i_S}$, and the j_Sth aspect, $\boldsymbol{D}_{j_S}$, respectively. Then, each column of $\boldsymbol{T}_{C_{i_S j_S}}$ should further be normalized by $d_{j_{n_j}} = \sum_{i=i_1}^{i_{n_i}} t_{i_{n_i} j_{n_j}}, j_{n_j} = 1, \cdots, i_{n_i}$. The normalized $\boldsymbol{T}_{C_{i_S j_S}}$ can thus be expressed as $\boldsymbol{T}_{C_{i_S j_S}}^{(N)} = \left[\frac{t_{i_u j_v}}{d_{j_v}}\right]_{i_{n_i} i_{n_j}}$. The normalized TRM, $\boldsymbol{T}_C^{(N)}$, can serve as the unweighted super-matrix

$\boldsymbol{W}$. To derive the weighted super-matrix, the values of the elements belonging to each submatrix, $\boldsymbol{T}_{C_{ij}}$, belonging to the matrix $\boldsymbol{T}_C$, can be added up and filled into a matrix $\boldsymbol{T}_D = \left[t_{c_{ij}}\right]_{m\times m}$, in which $t_{c_{ij}}$ is the sum of all the elements belonging to the submatrix $\boldsymbol{T}_{C_{ij}}$. Then, the matrix $\boldsymbol{T}_D$ can be normalized as $\boldsymbol{T}_D^{(N)} = \left[\frac{t_{c_{ij}}}{d_j}\right]_{m\times m}$ by normalizing each column to unity as follows, where $d_j = \sum_{i=1}^{m} t_{c_{ij}}$. The weighted super-matrix $\boldsymbol{\Pi}$ can be derived by multiplying the transposed $\boldsymbol{T}_D^{(N)}$ with $\boldsymbol{W}$, i.e., $\boldsymbol{\Pi} = \boldsymbol{T}_D^{(N)t}\boldsymbol{W}$. Then, the weighted super-matrix can be derived as $\lim_{\theta_e\to\infty} \boldsymbol{\Pi}^{\theta_e}$. Detailed explanations of the above process can further be found in [47]. The global priority vectors can be derived accordingly, along with the weights associated with each topic and aspect.

4. Empirical Study

This section presents a four-step procedure for social media mining and derivations of the criteria importance using the RF method, and the derivations of the influence relationships using the DEMATEL and the DANP. In this study, the psychological factors that can influence Taiwanese users' attitudes toward air pollution adaptation strategies were investigated. One of the major Taiwanese social media sites, the Dcard (dcard.tw), was mined to retrieve related posts. The topic modeling algorithm was then used to retrieve important topics from the social media data. After that, the topics were clustered according to their probability. The clusters were reviewed and then, based on the topics being associated with meaningful names, users' attitudes were assigned. Then, the feature importance of the topics was derived. Each topic served as the dependent variable in one analysis, while the rest of the topics served as the independent variables. The feature weights associated with the independent variables were derived. After normalization and transformation of these normalized feature weights into a five-point Likert scale, these feature weights served as the input for the DEMATEL as well as the DANP. The IRM and the influence weights were derived accordingly.

4.1. Scraping and Pre-Processing of Social Media Data

At first, Dcard (dcard.tw) a popular website with 4 million users that accounts for around one sixth of the weekly social media posts in Taiwan, was used to mine users' opinions regarding the air pollution problem in the country. Air pollution is one of the most serious and concerning environmental issues in emerging economies in general, and in Taiwan in particular. A total of 3700 messages related to air pollution were retrieved using the Application Programming Interface (API) of Dcard in September, 2020. However, some of these messages could be dated back to 2016. The posts were collected from a number of boards, including Mood, Chats, Science, News, Beauty, Life, etc. Since the posts being retrieved from Dcard were full of information unrelated to the analyses and included tremendous inconsistencies in the data, they were pre-processed and cleaned. After unrelated posts were removed, 1043 messages were left for further analyses. Punctuation, common stop words, infrequent words, duplicates, errors, and messages unrelated to air pollution were removed from the full texts using a program the authors coded in Python 3.7 [9].

4.2. Extracting the Main Topics Using the LDA methods

After the texts were cleaned, the LDA topic modeling method introduced in Section 3.1 was adopted to retrieve topics from the posts. The parameters were estimated after 1000 iterations of Gibbs sampling, using 12 topics for our data set. Based on the LDA, 12 topics with coherent groups of keywords (Table 1), which clearly described the associated meanings, were named by four environmental experts [9]. The 12 topics were fuel (t_1), masks (t_2), electronic cigarettes (e-cigarettes) (t_3), smoking (t_4), coal-fired power generation

(t_5), refuse combustion (t_6), power generation (t_7), policy ambiguity (t_8), climate change (t_9), wind power generation policy (t_{10}), allergies and health (t_{11}), and air purifiers (t_{12}).

Table 1. Identified topics and topic clustering.

No.	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
1	1	1	1	1	1	1	1	4	1	1	4	4
2	2	2	2	2	2	2	2	2	2	4	2	4
3	2	4	2	2	2	2	2	2	2	2	2	4
4	3	3	3	3	3	4	3	3	3	3	3	4
5	2	2	2	2	2	2	4	2	2	4	2	2
6	4	1	1	1	1	1	4	1	1	1	1	1
7	3	3	3	3	3	3	4	4	3	3	4	4
8	4	3	3	3	4	3	3	4	4	4	3	3
9	3	3	3	4	3	3	3	4	3	3	3	4
10	1	1	1	1	1	1	4	4	4	1	1	1
1035	1	2	1	1	1	1	1	1	1	4	4	4
1036	1	1	1	1	1	4	1	4	1	1	1	1
1037	2	2	2	2	2	4	2	2	2	2	4	4
1038	1	1	1	4	1	1	1	1	1	1	4	1
1039	3	3	3	3	3	4	3	3	4	3	4	3
1040	2	2	2	4	2	2	2	2	2	2	2	2
1041	1	1	1	1	3	1	1	4	4	1	1	1
1042	1	1	4	1	1	1	1	1	1	1	1	1
1043	3	3	3	3	3	3	3	3	3	3	4	4

Based on LDA, the per-document topic assignments z_{d_η}, and topic proportions θ_d are conducted. Each message (document) was assumed to have a mix of latent topics, and each topic was assumed to have a certain probability of occurring in the document. A document–topic matrix represented the relationship between document and topics. Each row in the matrix stood for a document and each column for a topic. An entry was the number of distribution probabilities of the document in the topic. The authors first normalized and standardized the document–topic matrix, and then used the quartile deviation to group the distribution probability. The lowest 25% of the document–topic matrix was defined as "1," the 25% to 50% portion was defined as "2," 50% to 75% defined as "3" and higher than 75% as "4" (see Table 1). The five highest probability terms in the top identified topics from the LDA topic modeling are summarized in Table 1. Then, the scales are normalized and transformed to Liker's 5-point scale for consistency with later methods.

4.3. Merging Similar Topic Using the Hierarchical Cluster Analysis

After the derivations of topics, the topics are classified further by using the hierarchical cluster analysis. Based on the results of cluster analysis, the topics were categorized into four clusters by using the SPSS statistical software (version 21.0), where the squared Euclidean distance was adopted to calculate dissimilarities between the clusters. (Refer [43] for the detailed analytic process.) Then, according to the features of the topics, the four clusters are labeled as egoistic concerns (EC), altruistic concerns (AC), biosphere concerns (BC), and adaptation strategies (AS), the four aspects of the value–belief–norm theory being proposed by Stern et al. [14] (refer Table 2).

Table 2. Five highest probability terms in the top identified topics from LDA topic modeling.

Cluster	Topic	Term/Importance				
		Term 1	**Term 2**	**Term 3**	**Term 4**	**Term 5**
Egoistic Concerns (EC)	Fuel (t_1)	U.S. 45.7	Taiwan 39.9	natural 39.5	fuel 34.3	smoking forbidden 25.9
	Mask (t_2)	air 193.3	air pollution 82.5	air quality 81.1	mask 74.9	research 67.3
	E-cigarette (t_3)	e-cigarette 504.3	tobacco 466.5	cigarette 190.1	Taiwan 140.3	harm reduction 122.3
	Smoking (t_4)	smokes 565.7	cigarette smoke 228.0	tobacco 129.2	smells 111.0	cigarette butts 75.1

Table 2. *Cont.*

Cluster	Topic	Term/Importance				
		Term 1	**Term 2**	**Term 3**	**Term 4**	**Term 5**
Altruistic Concerns (AC)	Coal-fired power (t_5)	Shen'ao power plant	air pollution	governmental	EPA (*)	coal burning
		96.9	93.4	56.9	56.1	47.2
	Refuse combustion (t_6)	air	garbage	earth	burning	joss paper
		53.9	44.5	39.5	32.4	26.2
	Power generation (t_7)	Tai-power	power plant	power unit	generator set	gas
		159.1	144.3	125.9	83.6	82.3
Biosphere Concerns (BC)	Policy ambiguity (t_8)	plebiscite	green with nuclear	nuclear	vote	government
		95.7	48.9	36.7	35.8	34.8
	Climate change (t_9)	climate	energy	global	climate change	renewable energy
		174.5	164.2	152.2	127.9	107.2
Adaptation Strategies (AS)	Wind power policy (t_{10})	Taiwan	wind power	offshore wind power	polar bear	offshore
		130.6	50.7	50.4	42.1	41.5
	Medical treatment (t_{11})	allergy	nose	pump	doctor	feel
		267.5	150.9	71.3	64.3	61.3
	Air purifier products (t_{12})	air purifier	allergy	recommend	air quality	air filter
		125.1	96.2	86.9	76.0	71.0

Note: * EPA is the abbreviation for the Environment Protection Agency, Taiwan.

4.4. Derivation of Feature Importance by Using the RM algorithm

Based on the results of topic modeling (see Table 1), for each topic, the feature importance of the other 11 topics was derived using the RandomForestRegressor in the Sci-Kit Learn Python toolkit [36]. For example, for the first topic (t_1), the feature importance of the other 11 topics was filled into the first column of the matrix $\boldsymbol{M}_F$ (see Table 3) by using Equation (5) in Section 3.2. For the second topic, (t_2), the feature importance of the other 11 topics was filled into the second column of the matrix. The same rule was applied to the rest of the topics. The largest element in each column was used to normalize the elements belonging to that column. Then, to be consistent with the definition of the IDRM of DEMATEL, the normalized result was multiplied by 5 to create the $\boldsymbol{\Omega}$ matrix using Equation (6) (Table 4 below). By calculating the average of the scores of the topics associated with any one post belonging to some specific aspect, the feature importnce matrix $\boldsymbol{M}_{F_a}$ and the $\boldsymbol{\Omega}_a$ of aspects could derived using the same approach by Equations (5) and (6). Since the aspect of biosphere concerns contained only two criteria, the RF and the DEMATEL were not applicable to most of the cases. Accordingly, the two topics belonging to the aspect of biosphere concerns were denoted as BC_1 and BC_2, respectively. These two matrices are demonstrated in Table 5 andTable 6 below.

Table 3. Feature Importance Matrix $\boldsymbol{M}_F$.

$\boldsymbol{M}_F =$													
	t_1	0.000	0.080	0.051	0.337	0.081	0.196	0.065	0.096	0.074	0.088	0.049	0.056
	t_2	0.040	0.000	0.060	0.053	0.057	0.043	0.098	0.030	0.076	0.057	0.051	0.049
	t_3	0.031	0.091	0.000	0.185	0.061	0.073	0.040	0.069	0.061	0.048	0.047	0.035
	t_4	0.446	0.114	0.428	0.000	0.034	0.039	0.053	0.048	0.057	0.050	0.184	0.060
	t_5	0.053	0.067	0.051	0.027	0.000	0.170	0.074	0.235	0.122	0.061	0.042	0.034
	t_6	0.095	0.076	0.140	0.032	0.203	0.000	0.073	0.141	0.082	0.138	0.039	0.044
	t_7	0.037	0.119	0.040	0.033	0.057	0.067	0.000	0.090	0.085	0.114	0.059	0.040
	t_8	0.056	0.116	0.054	0.032	0.286	0.214	0.228	0.000	0.222	0.118	0.054	0.060
	t_9	0.094	0.083	0.052	0.034	0.092	0.041	0.069	0.130	0.000	0.053	0.053	0.045
	t_{10}	0.056	0.101	0.036	0.038	0.035	0.082	0.178	0.067	0.081	0.000	0.062	0.136
	t_{11}	0.052	0.071	0.055	0.177	0.042	0.034	0.064	0.034	0.077	0.078	0.000	0.441
	t_{12}	0.040	0.081	0.032	0.051	0.052	0.041	0.059	0.058	0.063	0.197	0.360	0.000

Table 4. IDRM $\boldsymbol{\Omega}$.

$\boldsymbol{\Omega}$ =	t_1	0.000	3.366	0.594	5.000	1.407	4.581	1.416	2.044	1.665	2.250	0.682	0.639
	t_2	0.452	0.000	0.705	0.789	0.988	0.999	2.160	0.645	1.722	1.439	0.703	0.554
	t_3	0.352	3.839	0.000	2.740	1.057	1.708	0.870	1.474	1.363	1.211	0.647	0.394
	t_4	5.000	4.767	5.000	0.000	0.601	0.902	1.152	1.017	1.274	1.266	2.562	0.678
	t_5	0.594	2.827	0.595	0.400	0.000	3.976	1.622	5.000	2.746	1.549	0.579	0.388
	t_6	1.060	3.198	1.639	0.470	3.537	0.000	1.609	2.995	1.857	3.511	0.542	0.499
	t_7	0.412	5.000	0.472	0.486	0.991	1.561	0.000	1.917	1.924	2.889	0.818	0.450
	t_8	0.624	4.877	0.634	0.480	5.000	5.000	5.000	0.000	5.000	2.997	0.753	0.682
	t_9	1.057	3.476	0.602	0.502	1.607	0.955	1.508	2.770	0.000	1.344	0.741	0.507
	t_{10}	0.628	4.241	0.417	0.570	0.614	1.907	3.893	1.422	1.835	0.000	0.868	1.538
	t_{11}	0.578	2.979	0.647	2.622	0.741	0.783	1.412	0.726	1.735	1.978	0.000	5.000
	t_{12}	0.446	3.409	0.377	0.757	0.915	0.965	1.287	1.237	1.409	5.000	5.000	0.000

Table 5. Feature Importance Matrix $\boldsymbol{M}_{F_a}$.

$\boldsymbol{M}_{F_a}$ =	EC	0.000	0.421	0.145	0.203	0.420
	AC	0.395	0.000	0.613	0.471	0.190
	BC_1	0.456	0.359	0.000	0.108	0.360
	BC_2	0.058	0.151	0.075	0.000	0.030
	AS	0.092	0.069	0.167	0.217	0.000

Table 6. IRM $\boldsymbol{\Omega}_a$.

$\boldsymbol{\Omega}_a$ =	EC	0.000	5.000	1.179	2.153	5.000
	AC	4.328	0.000	5.000	5.000	2.261
	BC_1	5.000	4.259	0.000	1.151	4.281
	BC_2	0.633	1.786	0.609	0.000	0.358
	AS	1.007	0.818	1.365	2.305	0.000

4.5. Deriving the Influence Relationships/Weights Using DEMATEL and DANP

Based on the $\boldsymbol{\Omega}$ matrix being derived by the RF, the influence of topic i_d on topic j_d, denoted as $a_{i_d j_d}$ in the IDRM, will be equal to $\omega_{i_d j_d}$ in the i_dth row and the j_dth column. Thus, $\boldsymbol{A} = \boldsymbol{\Omega}$. By adopting the process introduced in Section 3.3, the TRM can be derived as shown in Table 7. Then, the row sum and column sum vectors of the TRM can be derived as $\boldsymbol{r}$ and $\boldsymbol{c}$ respectively in Table 8. The TRM of all the aspects as well as the $r_{i_d} + c_{i_d}$ and $r_{i_d} - c_{i_d}$ versus each aspect are demonstrated in Tables 9 and 10, respectively. The IRM is demonstrated in Figure 2. Further, the influence weights versus each topic and aspect can be derived according to the procedure outlined in Section 3.4. The results are demonstrated in Tables 8 and 10 respectively.

Table 7. The TRM of topics.

$\boldsymbol{T}_{\text{topics}}$ =	t_1	0.181	0.199	0.067	0.184	0.103	0.212	0.107	0.127	0.113	0.129	0.054	0.050
	t_2	0.033	0.194	0.039	0.044	0.059	0.068	0.101	0.059	0.091	0.082	0.038	0.036
	t_3	0.044	0.171	0.165	0.117	0.068	0.106	0.075	0.086	0.087	0.082	0.041	0.034
	t_4	0.180	0.242	0.182	0.195	0.072	0.105	0.099	0.096	0.100	0.105	0.099	0.060
	t_5	0.047	0.195	0.052	0.044	0.202	0.182	0.119	0.206	0.143	0.117	0.045	0.041
	t_6	0.054	0.172	0.073	0.045	0.157	0.204	0.106	0.133	0.107	0.156	0.042	0.043
	t_7	0.036	0.193	0.037	0.040	0.077	0.100	0.195	0.120	0.108	0.154	0.046	0.040
	t_8	0.062	0.291	0.060	0.057	0.225	0.241	0.240	0.235	0.237	0.195	0.061	0.066
	t_9	0.056	0.169	0.039	0.041	0.089	0.086	0.092	0.108	0.183	0.088	0.041	0.039
	t_{10}	0.043	0.188	0.037	0.045	0.066	0.115	0.168	0.091	0.103	0.197	0.053	0.071
	t_{11}	0.050	0.167	0.052	0.106	0.069	0.081	0.096	0.072	0.100	0.124	0.179	0.176
	t_{12}	0.048	0.179	0.042	0.072	0.071	0.097	0.107	0.091	0.099	0.212	0.180	0.180

Table 8. $r_{i_d} - c_{i_d}$, weight and ranking versus each topic.

	Topic	r_{i_d}	c_{i_d}	$r_{i_d} + c_{i_d}$	$r_{i_d} - c_{i_d}$	Weight	Rank
EC	t_1	1.527	0.832	2.359	0.694	9.948%	3
	t_2	0.842	2.362	3.204	−1.520	4.802%	12
	t_3	1.075	0.843	1.918	0.233	6.640%	10
	t_4	1.535	0.990	2.525	0.545	10.223%	2
AC	t_5	1.393	1.259	2.652	0.134	9.008%	5
	t_6	1.290	1.597	2.888	−0.307	8.336%	7
	t_7	1.145	1.504	2.650	−0.359	7.050%	9
BC_1	t_8	1.971	1.424	3.395	0.548	12.412%	1
BC_2	t_9	1.031	1.471	2.503	−0.440	6.413%	11
AS	t_{10}	1.178	1.642	2.820	−0.463	7.053%	8
	t_{11}	1.272	0.880	2.151	0.392	8.855%	6
	t_{12}	1.378	0.835	2.213	0.543	9.260%	4

Table 9. Total relation matrix $\boldsymbol{T}_{\text{dimensions}}$ of dimensions.

	EC	0.659	0.675	0.395	0.528	0.709
	AC	0.742	0.786	0.667	0.760	0.666
$\boldsymbol{T}_{\text{dimensions}}$ =	BC1	0.747	0.712	0.645	0.515	0.781
	BC2	0.163	0.223	0.142	0.416	0.155
	AS	0.214	0.207	0.201	0.307	0.454

Table 10. $r_{i_d} - c_{i_d}$ weight and ranking versus each aspect.

Symbol	r_{i_d}	c_{i_d}	$r_{i_d} + c_{i_d}$	$r_{i_d} - c_{i_d}$	Weight	Rank
EC	2.966	2.524	5.490	0.442	31.613%	1
AC	3.622	2.604	6.225	1.018	24.394%	3
BC_1	3.401	2.049	5.450	1.351	12.412%	4
BC_2	1.099	2.527	3.626	−1.428	6.413%	5
AS	1.383	2.766	4.149	−1.383	25.168%	2

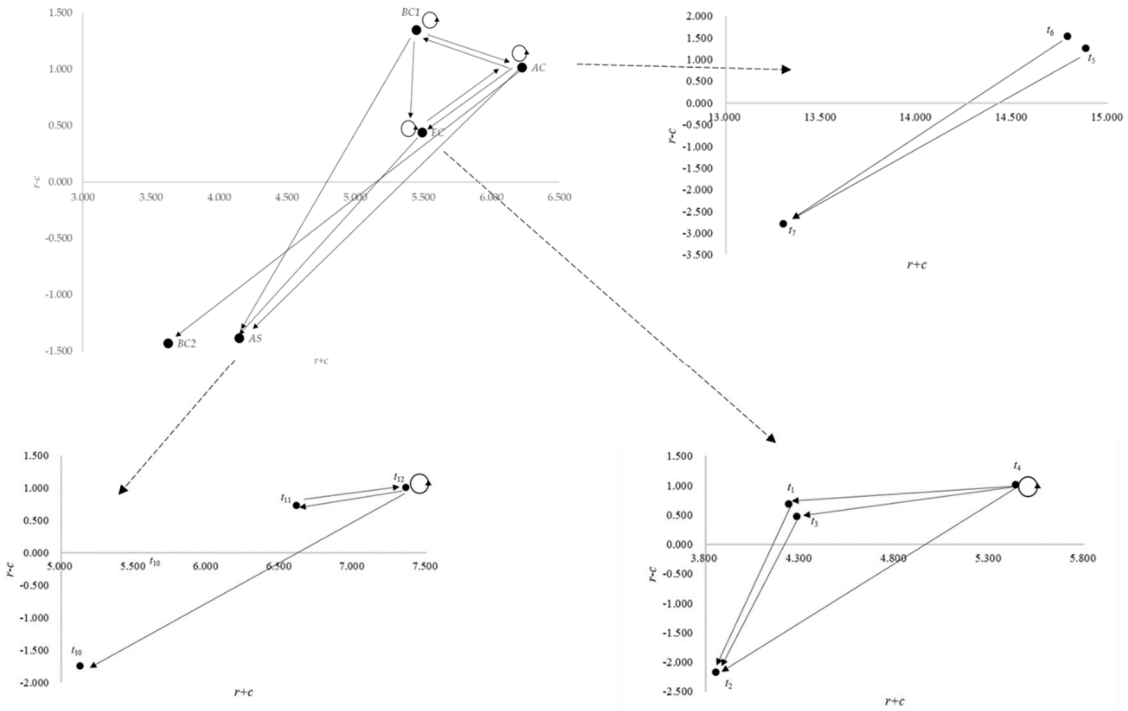

Figure 2. The IRM.

5. Discussion

In this work, a novel analytic framework, which consists of social media mining, RF, and MCDM techniques, was proposed. Further, the Taiwanese social media platform, Dcard, was used to retrieve data and validate the feasibility of the analytic framework. Meanwhile, influence relationships and influence weights were derived using the novel analytic framework. In the following section, the theoretical implications and advances in research methods presented in this study will be discussed.

5.1. Theoretical Implications

First, the mutual influence relationships among the three aspects from the VBN theory, i.e., altruistic, egoistic, and biosphere concerns, will be discussed. Based on the analytic results, the altruistic concerns influence both the egoistic and biosphere concerns. Furthermore, the biosphere concern influences the egoistic concern. The influence relationships are fully consistent with the original theoretical framework proposed by Stern et al. [14], which argues that the three environmental concerns—egoistic, altruistic, and biosphere—are mutually correlated. Environmental concern is the extent to which individuals are conscious of environmental issues and/or harms and support efforts to resolve those problems and/or point out an intention to contribute to the solution themselves [44]. According to Helm et al. [45], the three aspects are highly correlated. The less important influence relationships from egoistic concerns to biosphere concerns were not demonstrated in the IRM. This may be due to the lower value of total influence from egoistic concerns to the BC_1 aspect; thus, the influence was not demonstrated in Figure 2. The possible reason for this phenomenon may be the separation analysis of BC_1 and BC_2 aspects, which is limited by the infeasibility of deriving correct DEMATEL results based on the feature importance derived by using the RF algorithm, when there is only one dependent variable and one predictor. The unity feature importance derived will finally cause an IDRM with the same elements, for example, $[5]_{2 \times 2}$ in this case, where correct results cannot be derived by DEMATEL.

The influence relationships from egoistic concerns to adaptation strategies are consistent with past works. The adaptation strategy is a response strategy to environmental problems in general, and the air pollution problem in particular [46]. Adaptation strategies can provide possible adaptation plans/actions to facilitate the adjustment of human society and ecological systems to address environmental disasters by increasing a system's ability or reducing its vulnerability [51]. Effective adaptation strategies are vital for the long-term success of an organization [46]. Egoistic concerns are expressed as functional benefits and emotional benefits [52]. A person with egoistic concerns seeks individual economic benefits and emotional benefits [52]. Individuals with higher egoistic concerns will particularly think about the expenses and advantages of an environmental behavior for themselves [53]. Because air pollution is a local environmental problem that directly influences personal welfare, people may adopt adaptation strategies for individual benefit. According to the earlier work by the authors [9], egoistic concerns have significant correlations with adaptation strategies toward air pollution problems. When egoistic concerns are higher, more people are directly concerned with specific local environmental issues that directly impact them, rather than being stressed by global problems such as climate change [54]. We believe that people may adopt adaptation strategies for air pollution if air pollution problems are anticipated to influence the benefits of themselves. Based on the influence relationships being derived, i.e., EC→AS, people will adopt adaptation strategies such as supporting wind power generation policies (t_{10}), taking medical treatment (t_{11}), and purchasing air purifier products (t_{12}).

The influence relationships from altruistic concerns to adaptation strategies are also consistent with past works. Altruistic concern is a willingness to take action even in the face of the free rider problem [14], which means that individual self-interest is not sufficient to produce a collective good [55]. According to Stern et al. [14], although some people will possibly anticipate sufficient individual advantages or benefits to rationalize provision

of the collective good on egoistic grounds, most are also inspired by a more extensive, altruistic concern. Altruistic concern is a willingness to take action even in the face of the free rider problem [14], which means that individual self-interest is not sufficient to produce collective good [55]. Previous studies show that altruistic concerns may lead people to experience environmental stress and coping and then engage in pro-environmental activities [45]. Based on past works, altruistic concerns impact clients' purchase intentions regarding ecologically-friendly products [56]. According to the IRM in Figure 2, AC→AS, which means the influences from altruistic concerns are very important for the development of adaptation strategies. From the topics belonging to altruistic concerns, coal-fired power generation (t_5) and refuse combustion (t_6) are more important issues of concern to Taiwanese people. These air pollution-related problems influence consumer behavior toward purchasing air purifiers (t_{12}; 9.260%) and taking medical treatment (t_{11}; 8.855%). Though adopting wind power generation (t_{10}; 7.053%) is an alternative for reducing the threats caused by air pollution, the replacement of coal-fired or gas-fired power generation plants with green power needs long-term planning over many years. Therefore, wind power generation (t_{10}; 7.053%) is the least important strategy from Taiwanese social media users' perspective.

The influence relationship from biosphere concerns to adaptation strategies is also consistent with past works. Bio-spheric values reflect an individual's concerns/perception regarding the biosphere and highlight the quality of the natural environment, distinctly from its benefits to humans. Several studies have found that bio-spheric concerns are connected with pro-environmental behavior intention. According to Helm et al. [49], individuals with more bio-spheric concerns (for example, concern for living creatures and the environment) related to concerns about harmful impacts for all animals and plants on Earth might value the risks of climate change as more severe and stressful, and therefore will probably respond to them [57]. Thus, bio-spheric environmental concern is dominant in affecting psychological adaptation [45]. Nguyen et al. [58] pointed out that biosphere values stimulate active involvement in ecological consumption by enhancing clients' attitudes toward environmental protection and reducing problems related to environmentally-friendly products. Based on the work by Kiatkawsin et al. [59], bio-spheric values have more impact on customers' chances of purchasing sustainable merchandise. According to the IRM in Figure 2, the BC_1 (policy ambiguity) has more influence on the adaptation strategies than the BC_2 (climate change). The answer is very reasonable. First, based on the recognition of social media users, the influence of policy ambiguity (BC_1) is indeed stronger than that of climate change (BC_2). The terms associated with the only criterion (t_8) in BC_1, including the terms associated with the topic (green, nuclear, vote, government in Table 2), are those which have more influence on wind power generation policy (t_{10}). The stronger influence relationship can be observed from the TRM of topics in Table 7. The influence from t_8 to t_{10} (0.195) is indeed much higher than the influence from t_8 to t_{11} and t_{12}, which are 0.061 and 0.066, respectively. Further, the influence of climate change (t_9) on the three criteria in the AS aspect is 0.088, 0.041, 0.039, respectively. This means that policy ambiguity (BC_1) is indeed the major topic influencing the definition of wind power generation (AS).

Finally, according to the result of the DANP in Table 10, the influence weight for environmental concerns and adaptation strategies are prioritized as EC $\succ$ AS $\succ$ AC $\succ$ BC_1 $\succ$ BC_2. Many environmental issues are considered social dilemmas; that is, when individuals pursue their own self-interest, this results in damaging consequences for the collective. For example, Knes [60] proposed that promoting pro-environmental behavior is recognized as a moral issue by altruistic individuals but not by egoistic ones in the context of climate change. However, our study proposes that egoist concerns have a greater influence weight than altruistic and bio-spheric concerns in the context of air pollution. This may be why air pollution is one of the most pressing environmental and health issues, which can cause respiratory illnesses and allergies ranging from coughs to asthma, cancer, or emphysema. Related research by Vyver et al. [61] revealed that people who perceived

higher health threats were also more likely to engage in a range of pro-environmental behaviors in the case of turning off idling engines to reduce air pollution.

5.2. Advance in Research Method

The analytical framework which integrates the method of NLP, RF, and MCDM is a novel one which crosses the gap between social media mining and MCDM research. Numerous scholars have developed works using these methods individually. Very few scholars have tried to integrate the NLP methods with SEM. However, according to the authors' limited knowledge, this work is the first which tries to integrate these methods and derive meaningful results.

First, the RF algorithm can transform data retrieved from any database into the IDRM, which is required by DEMATEL. Traditionally, the MCDM method required opinions to be provided by experts. However, data retrieved from the database or the mass population (i.e., big data) can also provide very meaningful information. Thus, scholars have started to propose method(s) which tried to integrate the RF algorithm and the MCDM method, like DANP (e.g., the work by Liu et al. [62] and Lo et al. [63]), which provide insights into management problems based on real data. In this paper, the NLP-based social media mining techniques are further integrated and advance the existing RF and DANP-based method. Big data retrieved from social media can serve as the basis for uncovering social phenomena by using MCDM methods, which were difficult to achieve. However, the influence relationships can provide more meaningful information than traditional MCDM or statistical methods-based research.

Second, the social media mining-based MCDM framework can provide more insights into social phenomena or social theories. Traditionally, scholars used statistical sampling-based methods such as covariance-based SEM or PLS-SEM to verify the theoretical framework. The social media mining-based MCDM framework provides new opportunities for verifying causal relationships and deriving new influence relations and the importance of aspects belonging to the theoretical frameworks.

In general, the proposed analytical framework advances both the MCDM-based analytical framework and the methods for verifying social theories. The analytical framework can be further adopted in big data analytics, uncovering real problems and confirming social theories by using big data.

5.3. Limitations and Future Research Possibilities

From the aspect of limitations, the analytic results are derived based on the Taiwanese social media site. The results may be controversial when mining social media sites from other regions or economies. Meanwhile, the empirical results are based on the VBN theoretic framework. Whether the analytic framework can derive satisfactory results, which can be fully consistent with other social theories, is worth future study.

Further, as already mentioned in Section 5.1, when the number of criteria of some specific aspect is less than three, the RF based DANP may not be feasible. The unity feature importance will cause an IDRM with same elements, for example, $[5]_{2 \times 2}$. In this case, correct results cannot be derived by DEMATEL. Though this kind of situation will not really occur in research which refers to prior academic works, e.g., the confirmatory analyses based on SEM, which usually contain more than three to five criteria based on the questionnaires, the phenomenon actually constrains the development of some MCDM problems containing aspects with fewer than three criteria.

In the future, the novel analytic framework consisting of social media mining, RF, and MCDM methods can be used to retrieve more information from social media websites in general, and validate social theories regarding social phenomenon in particular. The newly derived influence relationships between altruistic and egoistic concerns and altruistic and biosphere concerns are also worth further research and investigation.

6. Conclusions

During the past decade, social media has emerged as one of the major sources for mining opinions from users in major and emerging economies. Though numerous scholars and practitioners have dedicated attention to mining useful information from social media, a lot more can be retrieved from the available data. The MCDM theories and methods have been well developed and widely applied to numerous economic, management, and engineering problems. However, very few scholars have tried to integrate the MCDM method with social media mining techniques. However, interesting results, such as influence relationships and valuable insights, can be retrieved from social media data. Thus, the authors proposed an analytic framework that integrates the LDA, RF, DEMATEL, and DANP. In this study, Dcard users' attitudes and adaptation strategies regarding air pollution problems were retrieved and analyzed based on the value–belief–norm theory proposed by Stern et al. [14].

Based on the analytic results, the influence relationships are fully consistent with the value-belief-norm theory. That is, altruistic concerns influence both egoistic and biosphere concerns. Furthermore, biosphere concerns influence egoistic concerns. Moreover, all three aspects—altruistic, egoistic, and biosphere concerns—influence adaptation strategies. The mutual influences between altruistic concerns and egoistic concerns, as well as altruistic concerns and biosphere concerns, were seldom discussed in past works. Whether these two influence loops are self-enhancing or self-attenuating is worth investigating further.

According to the results derived by the DANP, the most important aspects of the analytic framework include egoistic concerns and altruistic concerns, which had influence weights of 31.613% and 24.394%, respectively. The results are fully consistent with the authors' earlier work using the PLS-SEM to analyze the VBN theoretic framework [9], in which these two aspects were the ones most closely correlated with the adaptation strategies. That is, the influence relationships are consistent with statistical results.

The analytic results presented here were derived based on the Taiwanese social media site Dcard. The results may be controversial when mining social media sites from other regions or economies. Meanwhile, the empirical results were based on the VBN theoretical framework. Whether this analytic framework can derive satisfactory results that can be fully consistent with other social theories is a question worth further study. In the future, this novel analytic framework can be used to retrieve more information from social media websites in general, and validate social theories regarding social phenomenon in particular.

Author Contributions: C.-Y.H. designed, performed research, coded the random forest regression program, analyzed the data, wrote, and revised the paper. C.-L.Y. analyzed the data and wrote portions of the empirical study case. Y.-H.H. coded the data mining program. All authors have read and agreed to the published version of the manuscript.

Funding: This research was granted by MOST, Taiwan (MOST107-2629-M-492-001-MY2).

Institutional Review Board Statement: Not applicable. The study did not involve humans.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are not available because of ongoing studies.

Acknowledgments: The authors appreciate Yu-Sheng Kao for his initial discussion of the research ideas regarding to the analytic framework. Further we would thank Kao for his valuable opinion regarding to revising partial of the draft.

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

Appendix A. Notations and Abbreviations

Table A1. Notations.

Notations	Descriptions	Notations	Descriptions
$a_{i_d j_d}$	An element in matrix $\boldsymbol{A}$ of DEMATEL	$\boldsymbol{M}_F$	The feature importance matrix in RF
$\boldsymbol{A}$	Initial influence matrix of DEMATEL	$p_r(t)$	The ratio of training samples in t
B	The combinations of interaction terms of fixed size of possible interacting variables.	P	A joint probability distribution in RF
$\boldsymbol{c}$	Column sum vectors of the TRM in DEMATEL	p	Number of input variables of RF
d	Any document in the corpus of LDA	$\boldsymbol{r}$	Row sum vectors of the TRM in DEMATEL
D	A corpus in LDA	R	R means the subscript of the right node of t.
$\boldsymbol{D}_{i_S}$	The i_Sth aspect in DANP	s_t	A split in RF
$\boldsymbol{D}_{j_S}$	The j_Sth aspect in DANP	S	A training sample in RF
$H(y)$	$H(Y)$ the prior entropy of y	t	An internal node of a RF
I	Variable importances in RF	t_p	A topic in LDA
$\boldsymbol{I}_d$	Identity in DEMATEL	$t_{c_{ij}}$	An element of T_D
I_{mp}	Feature importance	T	The number of topics
i	Row index of $\boldsymbol{T}_D$	$\boldsymbol{T}_{\text{C}}$	The transposed matrix of the TRM
j	Column index of $\boldsymbol{T}_D$	$\boldsymbol{T}_D$	A matrix in DANP
i_d	i_dth row of IDRM	$\boldsymbol{T}_{C_{i_S j_S}}$	A submatrix of $\boldsymbol{T}_{\text{C}}$
j_d	j_dth column of IDRM	$\boldsymbol{T}_R$	Total relation matrix of DEMATEL
$i_p(t)$	impurity of node t in RF	T_{S}	A tree structure representing an input-output model
$\Delta i_p(t)$	The impurity reduction at node t	$v()$	$v(s_t)$ is the variable used in split s_t
i_F	Column index for the matrix $\boldsymbol{M}_F$	V	A set of categorical input variables of the RF
j_F	Row index for the matrix $\boldsymbol{M}_F$	V_R	All variables in V that are relevant to y.
i_{RF}	The subscript for the means $x_{i_{RF}}$ in RF	$\boldsymbol{W}$	The unweighted super-matrix
i_s	Row index for $\boldsymbol{T}_C$ in DANP	w_η	A word to be selected from φ_{t_p}
j_s	Column index for $\boldsymbol{T}_C$ in DANP	$x_1, \cdots, x_p$	Categorical inputs of the RF algorithm
i_u	Row index for the matrix $\boldsymbol{T}_{C_{i_S j_S}}$ in DANP	y	A categorical output in RF
j_v	Column index for the matrix $\boldsymbol{T}_{C_{i_S j_S}}$ in DANP	z_{d_η}	Per-document topic assignments
i_ω	Row index for the matrix $\boldsymbol{\Omega}$	z_η	A topic to be selected from θ_d
j_ω	Column index for the matrix $\boldsymbol{\Omega}$	α	The per-document topic distributions
k	Dimensionality of the Dirichlet distribution	β	The per-topic word distribution
k_r	The number of possible interacting variables in RF	φ_{t_p}	A multinomial distribution for a topic from a Dirichlet distribution
l_{j_f}	The largest element of the column of $\boldsymbol{M}_F$	φ_{z_η}	A multinomial distribution for the topic z_η
L	L means the subscript of the left node of node t.	θ_e	θ_e is the exponent of $\boldsymbol{\Pi}$
M	Number of documents in a corpus D in LDA	θ_d	θ_d is the topic distribution for document d
m	A subscript for an input (x_m) of RF	η	Index for the ηth word in LDA
m_s	Number of submatrices of DANP	ς	The exponent of N_R
n	Number of joint observations in RF	$\mathcal{P}_{k_r}(V^{-m})$	The set of subsets of V^{-m} of cardinality k_r
n_i	Number of topics in the i_Sth aspect in DANP	$\boldsymbol{\Omega}$	5 times the normalized result of $\boldsymbol{M}_F$
n_j	Number of topics in the j_Sth aspect in DANP	$\aleph$	A space where any t represents a subset of it
n_t	The number of training samples in t in RF	$\lvert\aleph_{i_{RF}}\rvert$	Number of sub-trees in RF
N_d	Number of words in a document in LDA	$\boldsymbol{\Pi}$	The weighted super-matrix in DANP
$\boldsymbol{N}_R$	The normalized IDRM of DEMATEL	ρ	A factor to normalize the IDRM
N_T	Number of trees in the forest of RF	$\omega_{i_\omega j_\omega}$	An element of the $\boldsymbol{\Omega}$ matrix

Table A2. Abbreviations.

Abbreviation	Definition	Abbreviation	Definition
AS	Adaptation strategies	IRM	Influence relation map
AC	Altruistic concerns	IDRM	Initial direct relation matrix
AHP	Analytic Hierarchy Process	IF-DEMATEL	Intuitionistic fuzzy DEMATEL
ANP	Analytic Network Process	LDA	Latent Dirichlet Allocation
API	Application programming interface	MCDM	Multiple-Criteria Decision-Making
BC	Biosphere concerns	NLP	Natural language processing
DDD	Data-driven Decision-Making	PLS-SEM	Partial least squares structural equation modeling
DEMATEL	Decision-Making Trial and Evaluation Laboratory	RF	Random forest
DANP	DEMATEL-based analytic network process	TOPSIS	Technique for Order Performance by Similarity to Ideal Solution

References

1. McCay-Peet, L.; Quan-Haase, A. What is social media and what questions can social media research help us answer. In *The SAGE Handbook of Social Media Research Methods*; Sloan, L., Quan-Haase, A., Eds.; Sage: London, UK, 2017; pp. 13–26.

2. Zafarani, R.; Abbasi, M.A.; Liu, H. *Social Media Mining: An Introduction*; Cambridge University Press: Cambridge, UK, 2014.
3. Fersini, E. Sentiment analysis in social networks: A machine learning perspective. In *Sentiment Analysis in Social Networks*; Pozzi, F.A., Fersini, E., Eds.; Morgan Kaufmann: Cambridge, MA, USA, 2017; pp. 91–111.
4. Jimenez-Marquez, J.L.; Gonzalez-Carrasco, I.; Lopez-Cuadrado, J.L.; Ruiz-Mezcua, B. Towards a big data framework for analyzing social media content. *Int. J. Inf. Manag.* **2019**, *44*, 1–12. [CrossRef]
5. Tan, W.; Blake, M.B.; Saleh, I.; Dustdar, S. Social-network-sourced big data analytics. *IEEE Int. Comput.* **2013**, *17*, 62–69. [CrossRef]
6. Lepri, B.; Staiano, J.; Sangokoya, D.; Letouzé, E.; Oliver, N. The tyranny of data? The bright and dark sides of data-driven decision-making for social good. In *Transparent Data Mining for Big and Small Data*; Cerquitelli, T., Quercia, D., Eds.; Springer: Cham, Switzerland, 2017; pp. 3–24.
7. Tang, J.; Chang, Y.; Liu, H. Mining social media with social theories: A survey. *ACM Sigkdd Explor. Newsl.* **2014**, *15*, 20–29. [CrossRef]
8. Provost, F.; Fawcett, T. Data science and its relationship to big data and data-driven decision making. *Big Data* **2013**, *1*, 51–59. [CrossRef] [PubMed]
9. Yang, C.-L.; Huang, C.-Y.; Hsiao, Y.-H. Using Social Media Mining and PLS-SEM to Examine the Causal Relationship between Public Environmental Concerns and Adaptation Strategies. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5270. [CrossRef]
10. Saaty, T.L. A scaling method for priorities in hierarchical structures. *J. Math. Psychol.* **1977**, *15*, 234–281. [CrossRef]
11. Liu, C.-H.; Tzeng, G.-H.; Lee, M.-H. Improving tourism policy implementation—The use of hybrid MCDM models. *Tour Manag.* **2012**, *33*, 413–426. [CrossRef]
12. Phillips-Wren, G.; Jain, L.C.; Nakamatsu, K.; Howlett, R.J. *Advances in Intelligent Decision Technologies: Proceedings of the Second Kes International Symposium Idt 2010*; Springer: Berlin, Germany, 2010.
13. Cheng, X.; Yan, X.; Lan, Y.; Guo, J. Btm: Topic modeling over short texts. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 2928–2941. [CrossRef]
14. Stern, P.C.; Dietz, T.; Abel, T.; Guagnano, G.A.; Kalof, L. A value-belief-norm theory of support for social movements: The case of environmentalism. *Hum. Ecol. Rev.* **1999**, *6*, 81–97.
15. Kaplan, A.M.; Haenlein, M. Users of the world, unite! The challenges and opportunities of Social Media. *Bus. Horiz.* **2010**, *53*, 59–68. [CrossRef]
16. Brynjolfsson, E.; McElheran, K. The rapid adoption of data-driven decision-making. *Am. Econ. Rev.* **2016**, *106*, 133–139. [CrossRef]
17. Baptista, J.; Wilson, A.D.; Galliers, R.D.; Bynghall, S. Social media and the emergence of reflexiveness as a new capability for open strategy. *Long Range Plan.* **2017**, *50*, 322–336. [CrossRef]
18. Kietzmann, J.H.; Hermkens, K.; McCarthy, I.P.; Silvestre, B.S. Social media? Get serious! Understanding the functional building blocks of social media. *Bus. Horiz.* **2011**, *54*, 241–251. [CrossRef]
19. Chauhan, P.; Sharma, N.; Sikka, G. The emergence of social media data and sentiment analysis in election prediction. *J. Ambient. Intell. Hum. Comput.* **2021**, *12*, 2601–2627. [CrossRef]
20. Fu, C.; Liu, W.; Chang, W. Data-driven multiple criteria decision making for diagnosis of thyroid cancer. *Ann. Oper. Res.* **2020**, *293*, 833–862. [CrossRef]
21. Saaty, T.L. *Decision Making with Dependence and Feedback: The Analytic Network Process*; RWS Publications: Pittsburgh, PA, USA, 1996.
22. Gabus, A.; Fontela, E. *World Problems, an Invitation to Further Thought within the Framework of DEMATEL*; Battelle Geneva Research Center: Geneva, Switzerland, 1972.
23. Yang, M.; Nazir, S.; Xu, Q.; Ali, S. Deep learning algorithms and multicriteria decision-making used in big data: A systematic literature review. *Complexity* **2020**, *2020*, 2836064.
24. Ouadah, A. Pipeline defects risk assessment using machine learning and analytical hierarchy process. In Proceedings of the 2018 International Conference on Applied Smart Systems (ICASS), Medea, Algeria, 24–25 November 2018; IEEE: Piscataway, NJ, USA, 2018.
25. Souissi, D.; Zouhri, L.; Hammami, S.; Msaddek, M.H.; Zghibi, A.; Dlala, M. GIS-based MCDM–AHP modeling for flood susceptibility mapping of arid areas, southeastern Tunisia. *Geocarto Int.* **2020**, *35*, 991–1017. [CrossRef]
26. Yasmin, M.; Tatoglu, E.; Kilic, H.S.; Zaim, S.; Delen, D. Big data analytics capabilities and firm performance: An integrated MCDM approach. *J. Bus. Res.* **2020**, *114*, 1–15. [CrossRef]
27. Muruganantham, A.; Gandhi, G.M. Framework for social media analytics based on multi-criteria decision making (MCDM) model. *Multimed. Tools. Appl.* **2020**, *79*, 3913–3927. [CrossRef]
28. Feldman, R.; Dagan, I. Knowledge Discovery in Textual Databases (KDT). In Proceedings of the KDD, Montreal, QC, Canada, 20–21 August 1995.
29. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv* **2017**, arXiv:1707.02919.
30. Trumbach, C.C.; Payne, D.; Kongthon, A. Technology mining for small firms: Knowledge prospecting for competitive advantage. *Technol. Forecast. Soc. Chang.* **2006**, *73*, 937–949. [CrossRef]
31. Demoulin, N.T.; Coussement, K. Acceptance of text-mining systems: The signaling role of information quality. *Inf. Manag.* **2020**, *57*, 103120. [CrossRef]
32. Kobayashi, V.B.; Mol, S.T.; Berkers, H.A.; Kismihók, G.; Den Hartog, D.N. Text mining in organizational research. *Organ. Res. Methods* **2018**, *21*, 733–765. [CrossRef] [PubMed]

33. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
34. Karami, A.; Lundy, M.; Webb, F.; Dwivedi, Y.K. Twitter and research: A systematic literature review through text mining. *IEEE Access* **2020**, *8*, 67698–67717. [CrossRef]
35. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211. [CrossRef]
36. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
37. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
38. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]
39. Louppe, G.; Wehenkel, L.; Sutera, A.; Geurts, P. Understanding variable importances in forests of randomized trees. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 431–439.
40. Yang, C.-L.; Huang, C.-Y.; Kao, Y.-S.; Tasi, Y.-L. Disaster Recovery Site Evaluations and Selections for Information Systems of Academic Big Data. *Eurasia J. Math. Sci. Technol. Educ.* **2017**, *13*, 4553–4589.
41. Huang, C.-Y.; Shyu, J.Z.; Tzeng, G.-H. Reconfiguring the innovation policy portfolios for Taiwan's SIP Mall industry. *Technovation* **2007**, *27*, 744–765. [CrossRef]
42. Tzeng, G.-H.; Huang, C.-Y. Combined DEMATEL technique with hybrid MCDM methods for creating the aspired intelligent global manufacturing & logistics systems. *Ann. Oper. Res.* **2012**, *197*, 159–190.
43. Yim, O.; Ramdeen, K.T. Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *Quant. Methods Psych.* **2015**, *11*, 8–21. [CrossRef]
44. Dunlap, R.E.; Jones, R.E. Environmental concern: Conceptual and measurement issues. In *Handbook of Environmental Sociology*; Greenwood Press: Westport, CN, USA, 2002.
45. Helm, S.V.; Pollitt, A.; Barnett, M.A.; Curran, M.A.; Craig, Z.R. Differentiating environmental concern in the context of psychological adaption to climate change. *Glob. Environ. Chang.* **2018**, *48*, 158–167. [CrossRef]
46. Laitinen, E.K. Long-term Success of Adaptation Strategies: Evidence from Finnish Companies. *Long Range Plann* **2000**, *33*, 805–830. [CrossRef]
47. Huang, C.-Y.; Chung, P.-H.; Shyu, J.Z.; Ho, Y.-H.; Wu, C.-H.; Lee, M.-C.; Wu, M.-J. Evaluation and selection of materials for particulate matter MEMS sensors by using hybrid MCDM methods. *Sustainability* **2018**, *10*, 3451. [CrossRef]
48. Huang, C.-Y.; Tung, I. Strategies for heterogeneous r&d alliances of in vitro diagnostics firms in rapidly catching-up economies. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3688.
49. Yang, C.-L.; Shieh, M.-C.; Huang, C.-Y.; Tung, C.-P. A derivation of factors influencing the successful integration of corporate volunteers into public flood disaster inquiry and notification systems. *Sustainability* **2018**, *10*, 1973. [CrossRef]
50. Tzeng, G.-H.; Huang, J.-J. *Multiple Attribute Decision Making: Methods and Application*; CRC Press: Boca Raton, FL, USA, 2011.
51. Mawdsley, J.R.; O'MALLEY, R.; Ojima, D.S. A review of climate-change adaptation strategies for wildlife management and biodiversity conservation. *Conserv. Biol.* **2009**, *23*, 1080–1089. [CrossRef]
52. Steg, L.; Bolderdijk, J.W.; Keizer, K.; Perlaviciute, G. An integrated framework for encouraging pro-environmental behaviour: The role of values, situational factors and goals. *J. Environ. Psychol.* **2014**, *38*, 104–115. [CrossRef]
53. De Groot, J.; Steg, L. General beliefs and the theory of planned behavior: The role of environmental concerns in the TPB. *J. Appl. Soc. Psychol.* **2007**, *37*, 1817–1836. [CrossRef]
54. Schultz, P.W. The structure of environmental concern: Concern for self, other people, and the biosphere. *J. Environ. Psychol.* **2001**, *21*, 327–339. [CrossRef]
55. Schwerin, D.S. Incomes policy in Norway: Second-best corporate institutions. *Polity* **1982**, *14*, 464–480. [CrossRef]
56. Prakash, G.; Choudhary, S.; Kumar, A.; Garza-Reyes, J.A.; Khan, S.A.R.; Panda, T.K. Do altruistic and egoistic values influence consumers' attitudes and purchase intentions towards eco-friendly packaged products? An empirical investigation. *J. Retail. Consum. Serv.* **2019**, *50*, 163–169. [CrossRef]
57. Schultz, P. Empathizing with nature: The effects of perspective taking on concern for environmental issues. *J. Soc. Issues* **2000**, *56*, 391–406. [CrossRef]
58. Nguyen, T.N.; Lobo, A.; Greenland, S. Pro-environmental purchase behaviour: The role of consumers' biospheric values. *J. Retail. Consum. Serv.* **2016**, *33*, 98–108. [CrossRef]
59. Kiatkawsin, K.; Han, H. Young travelers' intention to behave pro-environmentally: Merging the value-belief-norm theory and the expectancy theory. *Tour Manag.* **2017**, *59*, 76–88. [CrossRef]
60. Knez, I. Is climate change a moral issue? Effects of egoism and altruism on pro-environmental behavior. *Curr. Urban Stud.* **2016**, *4*, 157–174. [CrossRef]
61. Van de Vyver, J.; Abrams, D.; Hopthrow, T.; Purewal, K.; de Moura, G.R.; Meleady, R. Motivating the selfish to stop idling: Self-interest cues can improve environmentally relevant driver behaviour. *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *54*, 79–85. [CrossRef]
62. Liou, J.J.; Chuang, Y.-C.; Zavadskas, E.K.; Tzeng, G.-H. Data-driven hybrid multiple attribute decision-making model for green supplier evaluation and performance improvement. *J. Clean. Prod.* **2019**, *241*, 118321. [CrossRef]
63. Lo, H.-W.; Liou, J.J.; Huang, C.-N.; Chuang, Y.-C.; Tzeng, G.-H. A new soft computing approach for analyzing the influential relationships of critical infrastructures. *Int. J. Crit. Infrastruct. Prot.* **2020**, *28*, 100336. [CrossRef]