axioms

MDPI

Article

Monitoring and Recognizing Enterprise Public Opinion from High-Risk Users Based on User Portrait and Random Forest Algorithm

Tinggui Chen [1,*], Xiaohua Yin [1], Lijuan Peng [1], Jingtao Rong [1], Jianjun Yang [2] and Guodong Cong [3]

[1] School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China; yinxh0213@163.com (X.Y.); Cherrylijuanpeng@163.com (L.P.); rjt323@126.com (J.R.)
[2] Department of Computer Science and Information Systems, University of North Georgia, Oakwood, GA 30566, USA; Jianjun.Yang@ung.edu
[3] School of Tourism and Urban-Rural Planning, Zhejiang Gongshang University, Hangzhou 310018, China; cgd@mail.zjgsu.edu.cn
* Correspondence: ctgsimon@mail.zjgsu.edu.cn

Abstract: With the rapid development of "We media" technology, netizens can freely express their opinions regarding enterprise products on a network platform. Consequently, online public opinion about enterprises has become a prominent issue. Negative comments posted by some netizens may trigger negative public opinion, which can have a significant impact on an enterprise's image. From the perspective of helping enterprises deal with negative public opinion, this paper combines user portrait technology and a random forest algorithm to help enterprises identify high-risk users who have posted negative comments and thus may trigger negative public opinion. In this way, enterprises can monitor the public opinion of high-risk users to prevent negative public opinion events. Firstly, we crawled the information of users participating in discussions of product experience, and we constructed a portrait of enterprise public opinion users. Then, the characteristics of the portraits were quantified into indicators such as the user's activity, the user's influence, and the user's emotional tendency, and the indicators were sorted. According to the order of the indicators, the users were divided into high-risk, moderate-risk, and low-risk categories. Next, a supervised high-risk user identification model for this classification was established, based on a random forest algorithm. In turn, the trained random forest identifier can be used to predict whether the authors of newly published public opinion information are high-risk users. Finally, a back propagation neural network algorithm was used to identify users and compared with the results of model recognition in this paper. The results showed that the average recognition accuracy of the back propagation neural network is only 72.33%, while the average recognition accuracy of the model constructed in this paper is as high as 98.49%, which verifies the feasibility and accuracy of the proposed random forest recognition method.

Keywords: product user experience; enterprise network public opinion; identification of high-risk users; random forest algorithm; user portrait

Citation: Chen, T.; Yin, X.; Peng, L.; Rong, J.; Yang, J.; Cong, G. Monitoring and Recognizing Enterprise Public Opinion from High-Risk Users Based on User Portrait and Random Forest Algorithm. *Axioms* **2021**, *10*, 106. https://doi.org/10.3390/axioms10020106

Academic Editor: Davron Aslonqulovich Juraev

Received: 8 April 2021
Accepted: 21 May 2021
Published: 27 May 2021

1. Introduction

At present, with the popularization and development of media technology, a growing number of netizens are frequently posting opinions about enterprise products on social platforms, including some comments about poor experiences. However, this open transmission of information sometimes leads to comments about negative experiences, which are likely to cause negative public opinion. Some users may post negative comments via Weibo, for example, to compromise an enterprise, bringing huge damage to the business and resulting in a negative impact on business operations. However, due to the large number of users who post comments online, enterprises cannot monitor the public opinion

of all users. Therefore, personalized classification and identification of users can help enterprises develop more targeted solutions for public opinion and can also greatly reduce the cost of controlling public opinion for enterprises. Based on this, the monitoring and identification of users' online comments are particularly critical to improving the efficiency of managing public opinion and maintaining the corporate image.

Generally speaking, monitoring online public opinion for enterprises mainly refers to capturing the relevant public opinion information through a series of technical means to realize the monitoring and tracking of public opinion [1]. At present, the academic research on enterprises' online public opinion focuses on both macro and micro levels. With regard to research on the macro level, most scholars carry out a theoretical analysis on the communication characteristics of enterprise public opinion and propose suggestions to deal with negative public opinion. However, much of the macro research ignores the characteristics of the event content itself and is short of user tracking of negative public opinion events because such research is focused on the overall perspective. On the micro level, many scholars have introduced natural language processing tools to identify the emotional polarity of user comments and described the emotional distribution of product users. Although such studies are helpful for enterprises to understand the emotional tendencies of users, they lack personalized identification and monitoring of users in enterprises' negative public opinion events. In addition, in the study of recognition and detection algorithms, many scholars use neural network algorithms for classification and prediction, but this often requires manual classification of training samples in advance, and the recognition accuracy needs to be improved.

In this context, this paper combines user portrait technology and random forest algorithm to create a supervised model to recognize and monitor high-risk users expressing a public opinion on enterprises. The traditional technique of manually classifying training samples is abandoned in this model, while user portrait technology is adopted to classify the public opinion data of users. Further, such data are used as the training samples of the random forest algorithm, which lowers the workload of manual labeling and eliminates subjectivity, making it more scientific and objective. Specifically, Python (3.8 32-bit) is firstly used to crawl the information of users participating in the discussion of product experience. In turn, indicators are quantified according to the user's portrait technology. Consequently, the users are divided into high-risk, moderate-risk, and low-risk types in terms of the size of quantized indicators. Moreover, a high-risk user identification model based on a supervised random forest algorithm is established, which can learn and train the random forest classifier and predict whether newly released public opinion information users are high-risk users. Finally, by analyzing the public opinion of Hive Box and optimizing the parameters of the random forest algorithm, the parameters applicable to "Hive Box Charges" are selected to demonstrate the feasibility of the model. Compared with the BP (Back Propagation) neural network, the proposed model is verified by recognition results with high accuracy.

2. Literature Review

Negative public opinion caused by negative product experience has a significant impact on corporate image, and this has attracted the attention of many scholars. In this paper, user portrait technology and a random forest algorithm are used to monitor enterprise public opinion. Therefore, the applications of these two methods in enterprise public opinion are summarized.

Many scholars have studied the public opinion of enterprises from the perspective of user portraits. For example, Wang et al. [2] presented a novel learning model called the personal service ecosystem (PSE) to delineate user preferences and interests naturally. Virginia et al. [3] explored the relationship between the user's perceived quality, the club service dimensions, and the golf club performance. The results showed that the strategy to increase user satisfaction should be quite different depending on whether users were beginners or advanced golf players. Martínez-Cevallos et al. [4] analyzed the segmentation

of participants in a sports event according to their perceived quality, perceived value, satisfaction, and future intentions. Tiwari et al. [5] leveraged the power of the categorical information stored in the Wikipedia database to assign relative weights to entities that a user followed on Twitter. Zhang et al. [6] explored the application of user portrait technology in agriculture, and they constructed a situational recommendation system of agricultural science and technology resources based on user portraits. Sun and Chai [7] classified the portrait of online learners into three dimensions, and they constructed a tag system of learner portraits based on the data fields of an online learning platform. You et al. [8] excavated user characteristics through identified user behaviors and constructed user portraits based on behavior perception according to actual cases. Widiyaningtyas et al. [9] proposed a new similarity algorithm—so-called "user profile correlation-based similarity", which examined genre data and user profile data, namely, age, gender, occupation, and location. Ni et al. [10] designed a new user portrait construction architecture based on knowledge database construction and fingerprint matching technology. Chen et al. [11] modeled the public opinion reversal process, taking into consideration external intervention information and individual internal characteristics. Their simulation results showed that the intensity of external intervention information affected the direction and degree of public opinion reversal. Although the above literature works have promoted the development of user portrait theory, few scholars have applied this method to the research of enterprise network public opinion. Most scholars have only established an indicator system of user characteristics, while few scholars have quantified the indicators within it.

On the other hand, many scholars use random forests to study public opinion. For example, Yelkanat [12] investigated the performance of the random forest machine learning algorithm in estimating near-future case numbers for 190 countries in the world, and it was mapped in comparison with actual confirmed cases' results. Chen et al. [13] focused on studying the polarization phenomenon and established a model of public opinion dissemination with the polarization process considering individual heterogeneity. Simsekler et al. [14] used a tree-based machine learning algorithm as well as random forests to estimate accurate and stable associations. The results of our analysis showed that safety perception, management support, and supervisor/manager expectations were the leading drivers of patient safety grade. Chen et al. [15] introduced the social preference theory and revealed the micro-interaction mechanism of public opinion polarization. Different social preferences held by individuals had different influences on the public opinion polarization effects. Xiao et al. [16] held that the evolution of individual opinions was not only influenced by the interactions between neighboring individuals but was also updated naturally due to individual factors themselves in the absence of interactions. Li and Xiao [17] combined social judgment theory with the multi-agent model and proposed a multidimensional opinion evolution model for studying the dynamics of opinion polarization. The results demonstrated that polarization was influenced by the average degree of the network, and the polarization process was affected by the parameters of the assimilation effect and contrast effect. Although the above studies have promoted the development of the random forest algorithm, few scholars have applied it to the detection and identification of enterprise network public opinion users, and few scholars have combined it with user portrait technology.

This paper takes enterprise network public opinion as the research background. The current academic research in this field mainly focuses on the macro and micro perspectives. For research on the macro level, Regester and Larkin [18] hold that enterprise reputation is an important component of intangible assets, and they also put forward the 3T principle for enterprises to deal with a reputation crisis. Taylor and Perry [19] proposed that enterprises should establish a crisis response mechanism combining traditional media and network media when dealing with crises. By analyzing specific cases, Li and Dong [20] proposed the communication process model of enterprises' online public opinion and believe that the attention of netizens plays a leading role in the development direction of events. At the micro level, Yu [21] proposed the Leader Rank algorithm, which recognizes the emotional

characteristics of users by commenting on them and infers their positive, negative, or neutral attitude towards the author of an article. Yin et al. [22] examined how a perceived locus of the crisis was caused and used a real case study to perform a quantitative content analysis with a sample of 503 comments under online articles. Zhang and Zhang [23] used a sentiment analysis to calculate the emotional strength of defect events and explored the effectiveness of the trust repair strategy adopted by enterprises at different evolution stages. Bella et al. [24] selected the comment text of the official Twitter of an enterprise and used SVM (Support Vector Machines) to classify the text to construct an opinion leader style model. Liu et al. [25] used natural language processing technology to study the word segmentation of text information on a network and applied it for risk detection of corporate public opinion. Aggarwal and Singh [26] monitored user review data from different social network regions and optimized corporate advertising strategies. However, the described objects of the aforementioned research are products or services provided by enterprises, which lack the monitoring of users during negative public opinion events of enterprises.

3. Model Construction

After the outbreak of negative public opinions of an enterprise due to poor user experiences, monitoring high-risk users involves monitoring and recognizing those users participating in negative public opinion discussions. Firstly, user data in the network are crawled and classified by user portrait technology. Then, supervised random forest algorithms are adopted to identify and monitor high-risk groups. The overall framework of the paper is shown in Figure 1.

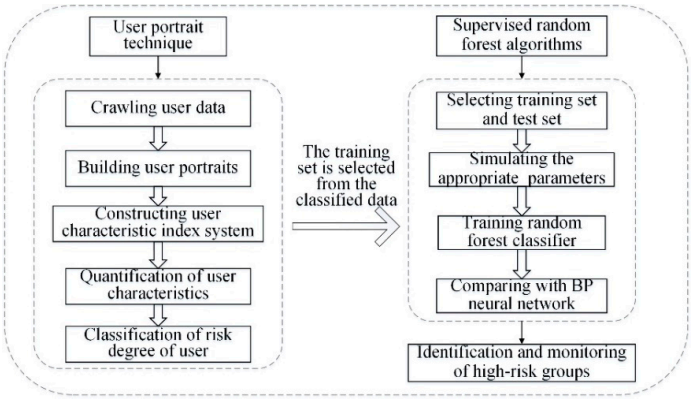

Figure 1. Research framework.

The variables and parameters involved in this paper are shown in Table 1.

Table 1. Relevant parameters.

Parameter	**Description**
Y	User's influence
Q	User's emotional tendency
S	User's base properties
H	User's activation
A	The number of thumbs up
B	The number of secondary comments
C	The number of fans
D	The amount of attention
F	Total number of Weibo posts shared
Gini	*Gini* coefficient
c	Sample category
nTree	The number of decision trees
maxf	The maximum number of features
leaf	The number of leaf nodes

3.1. Crawling User Data

The product name is used as a keyword to search on various social network platforms, and public opinion topics related to the product experience are selected. Subsequently, the Python (3.8 32-bit) crawler is used to crawl user information under the relevant topic, which is mainly for three kinds of information: (1) user information, such as the number of followers, address, and other account information; (2) user experience, comments, and other text information; and (3) interactive information, such as the number of reposts/comments, etc. In order to facilitate the subsequent emotional segmentation, pre-processing of the comment text data is required. Firstly, the symbols, expressions, and punctuations in the comments are removed. Secondly, comments that are not relevant to the company, such as advertisements and comments with very few words, are also deleted.

3.2. Building the User Portraits

A user portrait refers to the characterization of users through direct or indirect data, which is widely used in enterprise precision marketing, crime prevention, financial risk prediction, and other fields. In addition, by studying users' behaviors on the internet, user portraits can also be used for online public opinion governance [27]. At present, the user portrait system construction methods are varied, and user portrait construction methods with statistical analysis or rules are based on the practice of business knowledge combined with the understanding of the business, scenes, and problems. A characteristic index based on scenes is put forward to construct a user portrait framework and deal with real problems. In this paper, this method is adopted to construct an enterprise network public opinion user portrait. Based on the theoretical knowledge of enterprise public opinion, the user characteristic labels are crawled, and the data are quantified as indicators affecting the risk degree to build a user characteristic index system of enterprise public opinion and take it as the training data of the random forest.

3.2.1. Constructing the User Characteristic Index System

After the outbreak of negative public opinions on an enterprise, the user can participate in discussions and make comments on Weibo. While constructing the portrait of users participating in the discussions of negative public opinion, the characteristics can be divided into dynamic and static categories. The former is the attributes of users when they participate in social activities, including the time they participate in discussions and the opinions they hold. The latter describes basic properties, including an individual's ID, gender, site, number of followers, number of fans, number of Weibo posts shared, and other characteristics. Based on this, the specific user characteristic index system is shown in Table 2.

Table 2. User characteristic index system.

The First Index	The Second Index	The Third Index	The Fourth Index
User portrait	User's dynamic characteristics	User's emotional tendency Q	Comment text content
		User's influence Y	The number of thumbs up A The number of secondary comments B The number of fans C
	User's static characteristics	User's base properties S	Weibo ID Gender Site
		User's activation H	The amount of attention D Total number of Weibo posts shared F

3.2.2. Quantification of User Characteristics

In the index system constructed above, the static characteristics of users can be directly obtained according to the crawled Weibo user data. However, the dynamic characteristics of users change with the content posted by users, so it is necessary to quantify these characteristics.

For the quantification of emotional tendency, the Jieba segmentation tool [28] is adopted, and word frequency is counted. SnowNLP is used to score the emotion so as to obtain the emotional value of each posted text comment as a user's emotional tendency.

For the quantification of user influence, the influence of users in public opinion events is comprehensively judged according to the number of users' comments by likes, the number of secondary comments, and the number of fans. The calculation is shown in the following Formula (1):

$$Y_i = \frac{\frac{A_i}{A_{max}} + \frac{B_i}{B_{max}} + \frac{C_i}{C_{max}}}{3} \quad (1)$$

where Y_i represents the relative influence of the ith Weibo user; A_i represents the likes on a comment by user i, and A_{max} represents the maximum number of likes by all the users participating in the public opinion event; B_i represents the number of secondary comments of i, and B_{max} represents the highest number of secondary comments of all the users participating in the public opinion event; C_i represents the number of fans of i, and C_{max} represents the highest number of fans of all the users participating in the public opinion event.

With regard to the quantification of user activation, the participation in public opinion events can be comprehensively judged according to the amount of users' attention and the total number of posts on Weibo. The calculation is shown in the following Formula (2):

$$H_i = \frac{\frac{D_i}{D_{max}} + \frac{F_i}{F_{max}}}{2} \quad (2)$$

where H_i represents the relative activation of user i; D_i represents the number of followers of i, and D_{max} represents the highest number of followers of all users participating in the public opinion event; F_i represents the total number of Weibo posts shared by i, while F_{max} represents the maximum total number of Weibo posts shared by all users participating in the public opinion event.

3.2.3. Classification of Risk Degree of User

In our classification, those who post negative comments on Weibo and can easily bring great losses to a corporate image are defined as high-risk groups. On this basis, the risk levels of users who post comments on the internet are divided into three categories: high-risk users, medium-risk users, and low-risk users. According to the above-quantified user portrait index system, the risk degree of each user is positively correlated with the influence and activity of the user and negatively correlated with the affective tendency value of the user. Therefore, the user influence and user activity are sorted in descending order, while the emotional tendency is sorted in ascending order. High-risk users are defined as those who rank in the top 10% for influence, activity, and emotional propensity, and their label is set as 2. After the high-risk users are excluded from the top 20% of the users in the three indicators, the remaining users are defined as moderate-risk users, and the label is set as 1. After excluding high-risk and moderate-risk users, the remaining users shall be low-risk users, and the label shall be set as 0. The specific judgment criteria are shown in Table 3.

Table 3. Criteria for classification of user portraits.

Classification	Classification Label	User's Emotional Tendency Q	User's Influence Y	User's Activation H
High-risk type	2	$Q = 0$	$Y \in (0.1, Y_{max})$	$H \in (0.1, Y_{max})$
Moderate-risk type	1	$Q \in (0, 0.05)$	$Y \in (0.05, 0.1)$	$H \in (0.05, 0.1)$
Low-risk type	0	$Q \in (0.05, 1)$	$Y \in (0, 0.05)$	$H \in (0, 0.05)$

3.3. Supervised Random Forest Algorithms

After the outbreak of negative public opinion, monitoring and identifying high-risk users involves classifying those high-risk users participating in discussions of the product/experience, so as to identify the high-risk users with label 2. In recent years, the use of machine learning algorithms to classify all kinds of data has extended to a very wide range of applications, and there are various machine learning algorithms. Since the involved data belong to the supervised classification type, the random forest algorithm with strong generalization ability and fast training speed is adopted. The random forest algorithm was first proposed by Leo [29] in 2001, which was a classifier with a decision tree as the basic unit. It is equivalent to consisting of multiple decision tree classifiers. Each decision tree produces a result, and the final classification result is determined by voting. As such, the principle of random forest is introduced based on decision trees, and the specific steps of the algorithm are described as well.

3.3.1. Decision Tree

A decision tree algorithm is a supervised learning algorithm, which is mainly used for regression and classification. There are three optimal feature selection methods in the decision tree—namely, information gain, information gain rate, and Gini coefficient. By calculating the Gini coefficient selection characteristics, the user set $S = \{s_1, s_2, s_3, \ldots s_n\}$ is obtained, meaning that there are n different users participating in a public opinion discussion, and each user can be represented by $V = \{v_1, v_2, v_3, \ldots v_M\}$. It is assumed that each user in set S can be divided into k subsets, and each subset represents a class. The Gini coefficient [30] of the attribute v in dataset S is as follows:

$$Gini(S, v) \sum_{j=1}^{v} p(v_j) \times gini(S_j | V = v_j) \tag{3}$$

$$Gini(S_j | V = v_j) = \sum \sum_{i=1}^{n} p_i(1 - p_i) \tag{4}$$

In Formula (4), p_i is the probability that any user belongs to a specific category. By calculating the information gain of each feature, the feature with the lowest Gini coefficient in V is obtained, which is taken as the best feature.

3.3.2. Steps of Random Forest Algorithm

The random forest algorithm involves the integration of multiple decision tree classifiers. The tree classifier is set as an independent identically distributed random vector $\{h(x, \theta_k), k = 1, \ldots\}$. Given the dataset $T = \{(x_1, y_1), \ldots, (x_l, y_l)\}, l = 1, 2, \ldots, n\}$, the algorithm flow is as follows:

(1) Firstly, sampling with a replacement is adopted, and each sub-training set is extracted from n samples of the original training set by the replacement, namely $\{\theta_k\}$;

(2) Based on the decision tree, a binary tree corresponding to each sub-training set is formed. The process is as follows:

   (a) In the decision tree for constructing each node, select $m(M \geq m)$ features out of M as the candidate attribute characteristics for prediction or classification;

(b) Calculate the Gini coefficients of the selected m characteristics, and select the smallest attribute characteristic of the Gini coefficient as the optimal classification characteristic attribute;

(c) Classify the nodes according to the selected optimal attribute characteristics, and select the sub-characteristic attributes next to the optimal attribute characteristics from the remaining ones to ensure that each binary tree is a full binary tree.

(3) Repeat steps (1) and (2) until the generated tree can accurately classify the samples in the training set and combine all tree models in the operation process to form a random forest model;

(4) For any test sample, the final classification result of the sample is usually decided by simple voting. The specific formula is as follows:

$$c = argmax_c\left(\frac{1}{ntree}\sum_{k=1}^{ntree} I(h(x,\{\theta_k\}) = c)\right) \quad (5)$$

where $I(\cdot)$ is the indicator function, and c is the sample category that receives the most votes.

The specific algorithm diagram is shown in Figure 2.

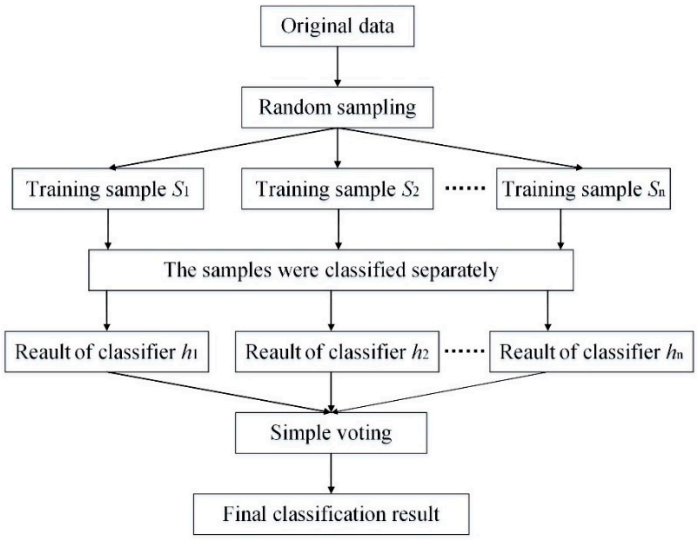

Figure 2. Random forest algorithm diagram.

3.4. Combination of the User Portrait and Random Forest Algorithm

Based on the theory of user portraits, first, the data of user characteristics are crawled, and the user characteristic index system is built up. Then, the quantitative characteristics after the indicators are considered as independent variables, and the user's degree of risk is considered as the dependent variable. According to the characteristics of the target, the degrees of risk to the user classification are sorted out. Furthermore, data from the classification of the dataset are selected as a training set for the random forest algorithm through continuously generating a decision tree.

4. The Empirical Analysis

In this section, Hive Box (one of the largest express enterprises in China), a typical representative of the logistics industry, is selected as the research object to conduct an empirical analysis of the recent outbreak of a "Hive Box charges" public opinion event.

On 30 April 2020, Hive Box launched a membership system in which ordinary users could keep a package for free within 12 h, for CNY 0.5 for 12 h after overtime, and for CNY 3 for capping. After the implementation of this system, most users' experiences of Hive

Box were very poor, and they made negative comments on various social platforms. On 8 May, some communities in Hangzhou and Shanghai posted notices to stop using Hive Box, and the negative public opinion continued to spread online. As of 11 May, the number of discussions about Hive Box on Weibo had reached 3.5 million. In view of the negative public opinion of Hive Box caused by poor user experiences, identification and monitoring of the online public opinion of the aforementioned enterprise were applied to depict the user portrait characteristics of Hive Box enterprise and to identify high-risk groups, so as to take targeted measures to improve user experience.

4.1. Crawling User Data

"Hive Box" and "Hive Box charges" were chosen as the main key words, the related Weibo comments starting from "30 April 2020" were selected, and the Python (3.8 32-bit) software crawl was adopted to collect user data—mainly user information, comment information for experience, and interactive information. In total, 48,267 data pieces were processed by Python (3.8 32-bit); a piece of data refers to a single comment posted by a user on the network, and it can include many bytes of varying lengths. As such, about 44,775 valid pieces of data were obtained after removing useless punctuations and emojis.

4.2. Building the User Portrait

Firstly, the characteristics of the users were quantified according to the above-built model. User comment text was segmented using the Jieba tool and was scored through the natural language processing tool SnowNLP. The text emotional value Q was quantified and mapped to [0,1] as the user's emotional tendency. The specific distribution is shown in Figure 3.

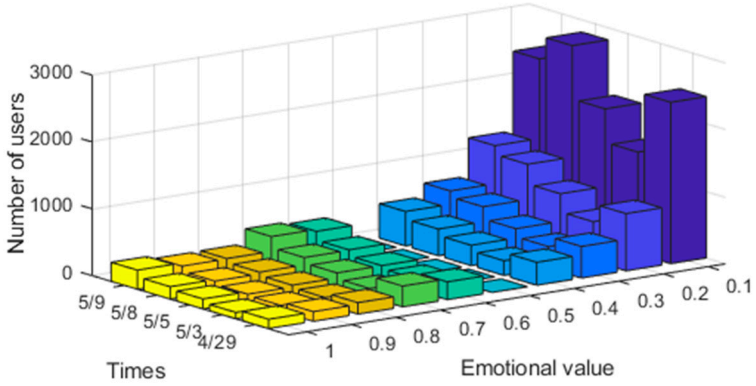

Figure 3. Distribution of emotional value for "Hive Box charges".

Figure 3 shows that in the initial stage of the "Hive Box charges", most users' emotional tendency was less than 0.1, a few users' emotional value was 0.5, and a small number of users' emotional value was more than 0.5. According to the calculation, 86.19% of users' emotional value was less than 0.5, which shows that most users held a negative emotional tendency towards this event, while only a small number of users held a positive attitude, and almost no users held a neutral attitude. Subsequently, on 8 May, a number of communities in Hangzhou and Shanghai jointly boycotted Hive Box, and most of the negative emotions of users also reached a peak. It can be seen that this event had a great impact on the corporate image, and the user experience was very poor. In such a case, Hive Box enterprise would need to take immediate strategies to improve the user experience.

4.3. Classification of Risk Degree of User

With regard to the quantification of influence and activation, the influence $Y \in (0, 0.666709)$ and the activation $H \in (0, 0.83015)$ were calculated using the above Formulas (1) and (2), and then, the users participating in the public opinion discussion were classified according to the above classification criteria. The final results are shown in Table 4.

Table 4. Classification of high-risk groups.

User Type	High-Risk	Moderate-Risk	Low-Risk
User number	26	151	44,598
Proportion	0.058%	0.34%	99.6%

Table 4 shows that most users were low-risk, and only a small number were high-risk, accounting for only 0.058% ratio, which indicates that although the majority of users held a negative opinion, most of them were ordinary users in the network with limited influence and were not significant enough to become high-risk users. Based on this, in order to prevent the continuous spread of negative public opinions, Hive Box enterprise could focus on monitoring the high-risk users and improving user experience.

4.4. Selecting Training Set and Test Data for Random Forest Algorithm

According to the classification of data, influence Y, activation H, and emotional value Q were deemed as attributes, and a training set was randomly selected from the data classified using the user portrait technique. A random forest classifier was trained using the random forest algorithm. From the data excluding the training samples, 1000 pieces of data were randomly selected as the test sample to identify the corresponding category of each sample, and the recognition accuracy of the algorithm was calculated. The training set and test set were disjoint.

4.5. Simulating the Appropriate Parameters

As different parameters have a great influence on the random forest algorithm's identification precision, different datasets generally use different values of parameters. Selecting appropriate parameters not only improves the identification precision of the algorithm but also speeds up the training. Therefore, four parameters in the random forest algorithm were simulated and analyzed: the number of decision trees (*nTree*), the size of the training sample, the maximum number of features (*maxf*), and the number of leaf nodes (*leaf*). The suitable parameters for this case were determined by simulation analysis. The specific results are shown in Figures 4 and 5.

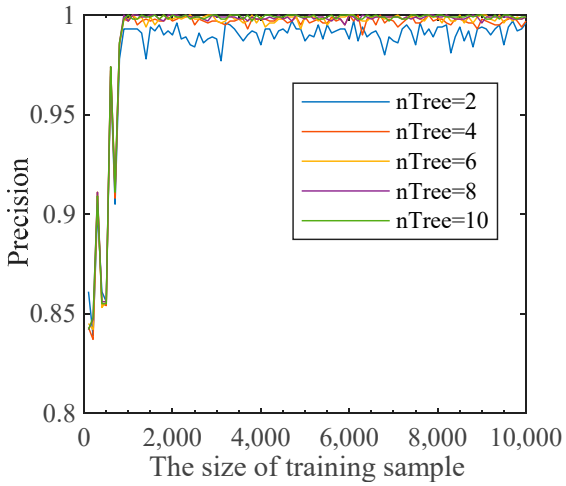

Figure 4. Simulation of *nTree* and the size of training sample.

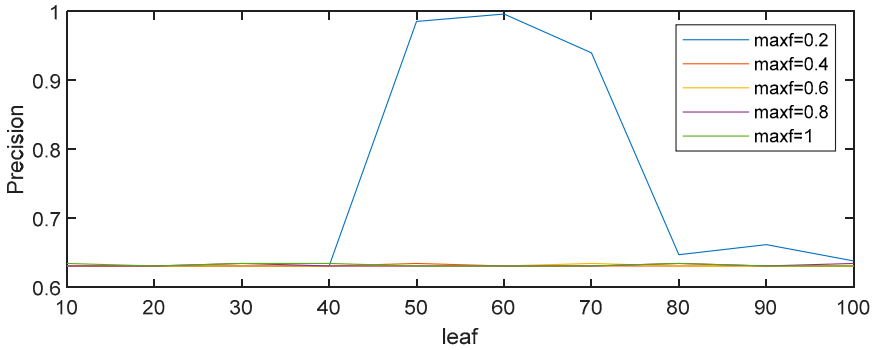

Figure 5. Simulation of *leaf* and *maxf*.

As can be seen from Figure 4, when the training sample is small, the accuracy of recognition of the high-risk population fluctuates greatly. However, when the training sample is larger than 1000, the recognition accuracy fluctuates only slightly. Therefore, in reality, the high-risk population monitoring and identification model requires at least 1000 pieces of data. Since there are enough data crawled in this paper, 10,000 pieces of data were selected as the training sample. In addition, with the increase in the number of decision trees in the random forest, the recognition accuracy is further improved, but at the same time, the running speed needs to be taken into account. Therefore, *nTree* = 10 was selected for training. As shown in Figure 5, only when *maxf* = 0.2 will the prediction accuracy change, and when *leaf* =60, the identification accuracy will be the highest. Therefore, in this case, our simulation set the parameters *maxf* = 0.2 and *leaf* = 60. Based on the above simulation results, the parameter settings are shown in Table 5 below.

Table 5. Parameter settings.

Parameter	The Size of Training Sample	*nTree*	*Maxf*	*Leaf*
setvalue	>1000	10	0.2	60

4.6. Training Random Forest Classifier and Comparing the Results with BP Neural Network

The trained classifier was used to identify the test samples. In order to improve the reliability of the results, 2000 pieces of data were randomly selected as the test set and divided into two groups. BP neural network and random forest were used to identify the samples, respectively, and the identification accuracy is shown in Table 6.

Table 6. Comparison of identification accuracy.

Test Dataset	Recognition Accuracy of BP Neural Network	Recognition Accuracy of Random Forest
Dataset 1	62%	98.98%
Dataset 2	84%	97.8%
Dataset 3	71%	98.7%
Average identification accuracy	72.33%	98.49%

In Table 6, the identification accuracy of random forest in the three datasets is much higher than that of BP neural network. The average identification accuracy of random forest is 98.49%, while the average accuracy of BP neural network is only 72.33%. Since 1000 pieces of data are too large an amount, in order to better visualize the identification results, 100 pieces of data were selected from each dataset for display. The results are shown in Figures 6–11.

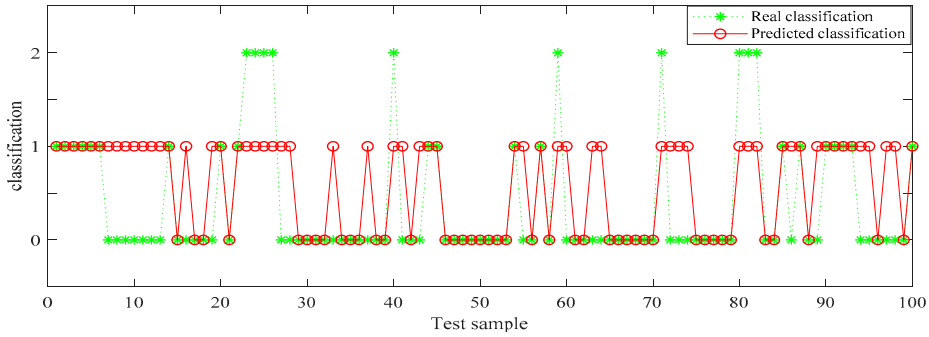

Figure 6. BP (back propagation) neural network recognition results from dataset 1.

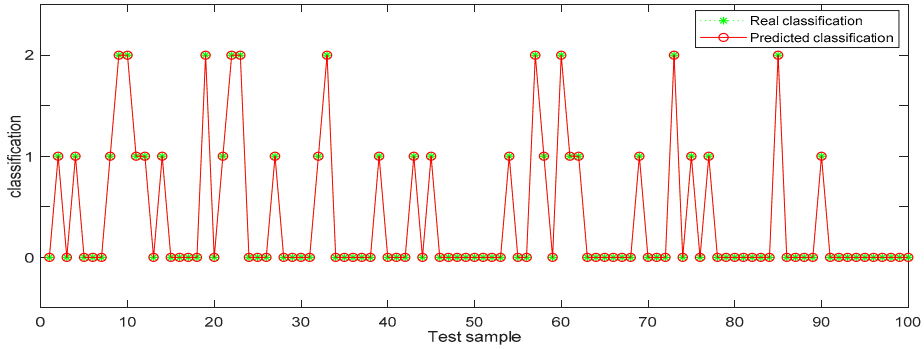

Figure 7. Random forest identification results from dataset 1.

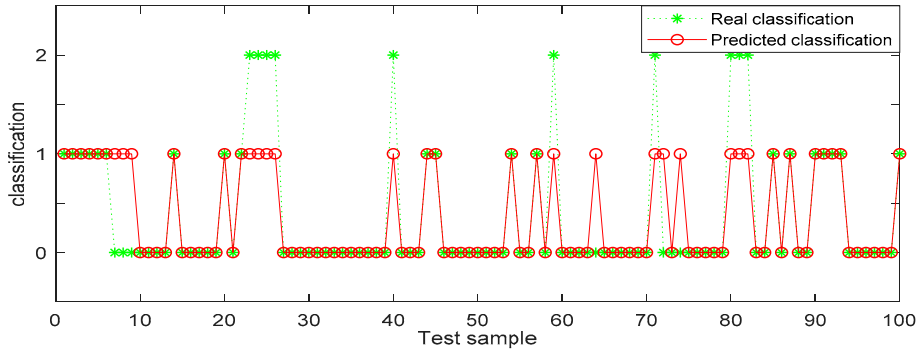

Figure 8. BP (back propagation) neural network recognition results from dataset 2.

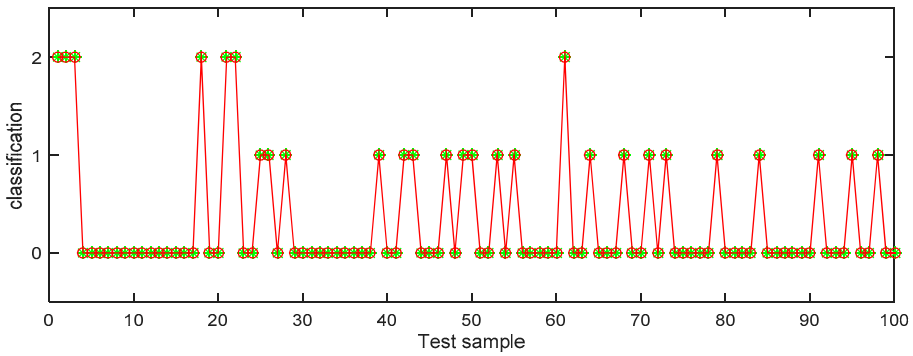

Figure 9. Random forest identification results from dataset 2.

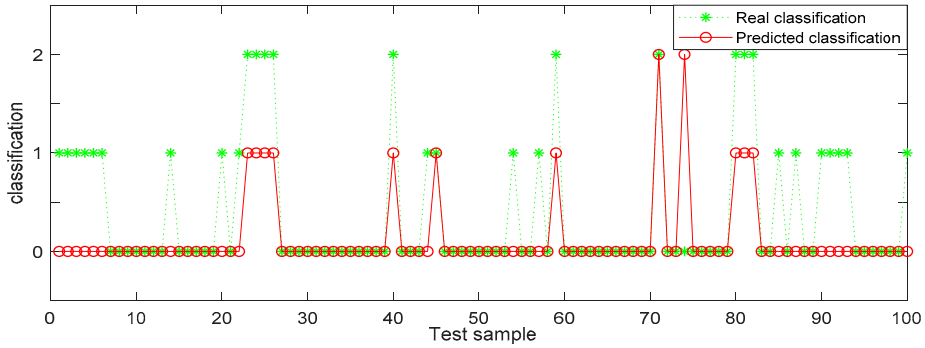

Figure 10. BP (back propagation) neural network recognition results from dataset 3.

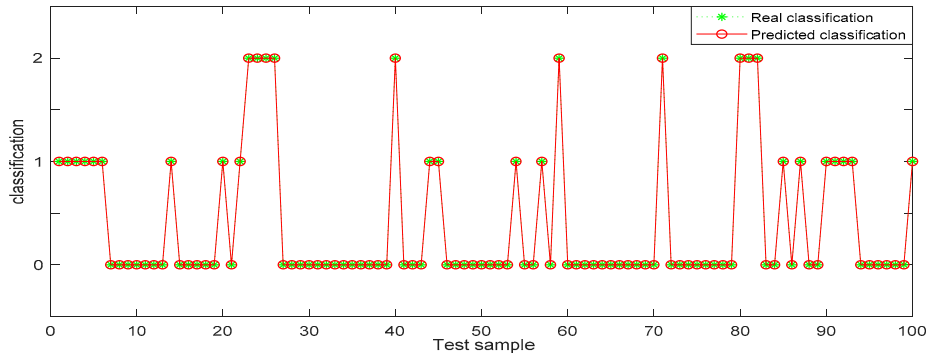

Figure 11. Random forest identification results from dataset 3.

In the three datasets, most of the users who made comments are low-risk users, while only a small number of users are high-risk users, which is also consistent with the data characteristics in Section 4.3 above. In the 100 test samples of Figures 6–11, the category of users identified by random forest is consistent with the real category, but the recognition results from BP (back propagation) neural network are only partially accurate. This suggests that the random forest identification method has higher accuracy and feasibility, which can be used for identifying high-risk users in public opinion discussions of an enterprise network.

4.7. Discussion

User portrait technology and the random forest algorithm were used to monitor and identify the public opinion events related to "Hive Box charges". The related discussions are as follows:

(1) When the sample size is big enough, the results of training indicate a dependence of precision on the training sample size. However, if the sample size is small, the result of training will be influenced by the size of the sample. This may be due to contingency. Therefore, we tested different sample sizes and demonstrated that the required sample size is at least above 1000 in the "Hive Box charges" public opinion event, and the identification accuracy will not be affected by the sample size;

(2) With regard to the "Hive Box charges" public opinion event, although 86.19% of the users participating in the discussion held negative emotions, most of them were low-risk users and only 0.058% were high-risk users. Therefore, monitoring the public opinion of high-risk groups could greatly save the costs of public opinion control of the enterprise;

(3) In the random forest algorithm, the final result is the average output result of each decision tree. Generally speaking, the larger the number of decision trees is, the stronger the robustness and the higher the accuracy of the random forest algorithm are. However, if the number of decision trees is too large, it is easy to increase

the correlation between the trees and increase the running time of the algorithm. Therefore, determining the appropriate number of decision trees is helpful to improve accuracy and efficiency. Our simulation showed that in the "Hive Box charges" public opinion event, the appropriate number of decision trees is 10;

(4) Compared with the BP neural network algorithm, the random forest algorithm used in this paper has a higher identification accuracy and is better for identifying high-risk groups of enterprise public opinion.

5. Conclusions

From the perspective of improving a user's experience with an enterprise, this paper proposes a monitoring and identification method combining user portrait technology and random forest algorithm to find high-risk users holding a negative enterprise public opinion. Accordingly, the traditional way of manual classification of training samples is abandoned. The proposed scheme can help enterprises to identify high-risk users that have had a poor experience and who may trigger negative public opinion. The main contributions of this paper are as follows:

(1) The traditional supervised machine learning algorithm is abandoned, while user portrait technology is adopted to classify the public opinion data of enterprise users. Furthermore, such data are used as the input data for the random forest algorithm, which lowers the workload of manual labeling and is not subjective, making it more scientific and objective;

(2) Combining the user portrait technique and the random forest algorithm, a model to identify public opinions on enterprises from high-risk users in terms of user experience of a product was established, which could allow enterprises to identify and monitor high-risk groups that may threaten their network image. The user portrait technique can identify the characteristics of users, providing a new perspective for the management of enterprise public opinion;

(3) In analyzing the public opinion of Hive Box, after optimizing the parameters of the random forest algorithm, the parameters applicable to "Hive Box charges" were selected to demonstrate the feasibility of the model. Compared with the BP neural network recognition results, the proposed model was verified by the recognition results with high accuracy.

However, there are still some limitations in this paper that need to be further explored. In this paper, only users who participated in the public discussion of opinions and experiences were identified, and no specific countermeasures were proposed. Therefore, further research could carry out a specific analysis based on the comments of high-risk users' experience perception to help enterprises come up with targeted measures.

Author Contributions: T.C. described the proposed framework and wrote the whole manuscript; X.Y. implemented the simulation experiments; L.P. and J.R. collected data; J.Y. and G.C. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Social Science Foundation of China (Grant No. 20BTQ059), the China (Hangzhou) cross-border electricity business school, Contemporary Business and Trade Research Center and Center for Collaborative Innovation Studies of Modern Business of Zhejiang Gongshang University of China (Grant No. 14SMXY05YB), Zhejiang Federation of Humanities and Social Sciences funded project, China (Grant No. 2019N21), Research Topics Project in higher education of Zhejiang Gongshang University (Grant No. Xgy20034), Discipline Construction and Management Project of Zhejiang Gongshang University (Grant No. XXK2019007), as well as First Class Discipline of Zhejiang-A (Zhejiang Gongshang University-Statistics).

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Zhang, J. Research summary of Network public opinion information mining in China. *Intell. Sci.* **2016**, *34*, 167–172.
2. Wang, H.; Tu, Z.; Fu, Y.; Wang, Z.; Xu, X. Time-aware user profiling from personal service ecosystem. *Neural Comput. Appl.* **2020**, *33*, 3597–3619. [CrossRef]
3. Serrano-Gómez, V.; García-García, Ó.; Pinasa, I.V.G.; Fernández-Liporace, M.; Hernández-Mendo, A.; Rial-Boubeta, A. Measuring perceived service quality and its impact on golf courses performance according to types of facilities and user profile. *Sustainability* **2020**, *12*, 5746. [CrossRef]
4. Martínez-Cevallos, D.; Proao-Grijalva, A.; Alguacil, M.; Duclos-Bastias, D.; Parra-Camacho, D. Segmentation of participants in a sports event using cluster analysis. *Sustainability* **2020**, *12*, 5641. [CrossRef]
5. Tiwari, S.; Saini, A.; Paliwal, V.; Singh, A.; Mattoo, R. Implicit preferences discovery for biography recommender system using twitter. *Procedia Comput. Sci.* **2020**, *167*, 1411–1420. [CrossRef]
6. Zhang, H.; Qin, X.; Zheng, H. Research on contextual recommendation system of agricultural science and technology resource based on user portrait. *J. Phys. Conf. Ser.* **2020**, *1693*, 012186. [CrossRef]
7. Sun, Y.; Chai, R. An early-warning model for online learners based on user portrait. *Ingénierie Systèmesd' Inf.* **2020**, *25*, 535–541. [CrossRef]
8. You, M.; Yin, Y.; Lu, S. User portrait based on behavior perception. *J. Zhejiang Univ.* **2021**, *4*, 1–8.
9. Widiyaningtyas, T.; Hidayah, I.; Adji, T.B. User profile correlation-based similarity algorithm in movie recommendation system. *J. Big Data* **2021**, *8*, 52. [CrossRef]
10. Ni, J.; Li, W.; Dong, W.; Zhuang, Y. User behavior detection and portrait construction system based on web log. *Comput. Era* **2021**, *11*, 42–46.
11. Chen, T.; Wang, Y.; Yang, J.; Cong, G. Modeling multidimensional public opinion polarization process under the context of derived topics. *Int. J. Environ. Res. Public Health* **2021**, *18*, 472. [CrossRef]
12. Yelkanat, C.M. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. *Chaos Solitons Fractals* **2020**, *140*, 110210. [CrossRef]
13. Chen, T.; Rong, J.; Yang, J.; Cong, G.; Li, G. Combining public opinion dissemination with polarization process considering individual heterogeneity. *Healthcare* **2021**, *9*, 176. [CrossRef] [PubMed]
14. Simsekler, M.C.E.; Qazi, A.; Alalami, M.; Ellahham, S.; Ozonoff, A. Evaluation of patient safety culture using a random forest algorithm. *Reliab. Eng. Syst. Saf.* **2020**, *204*, 107186. [CrossRef]
15. Chen, T.; Li, Q.; Fu, P.; Yang, J.; Xu, C.; Cong, G.; Li, G. Public opinion polarization by individual revenue from the social preference theory. *Int. J. Environ. Res. Public Health* **2020**, *17*, 946. [CrossRef]
16. Xiao, R.; Yu, T.; Hou, J. Modeling and simulation of opinion natural reversal dynamics with opinion leader based on HK bounded confidence model. *Complexity* **2020**, *2020*, 7360302. [CrossRef]
17. Li, J.; Xiao, R. Agent-based modelling approach for multidimensional opinion polarization in collective behaviour. *J. Artif. Soc. Soc. Simul.* **2017**, *20*, 14. [CrossRef]
18. Regester, M.; Larkin, J. Risk issues and crisis management in public relations: A casebook of best practice. *Res. Nurs. Health* **2008**, *5*, 93–101.
19. Taylor, M.; Perry, D.C. Diffusion of traditional and new media tactics in crisis communication. *Public Relat. Rev.* **2005**, *31*, 209–217. [CrossRef]
20. Li, G.; Dong, Q. Research and empirical analysis on the communication process of enterprises' online public opinion under Web2.0 environment. *Intell. Sci.* **2011**, *29*, 1810–1814.
21. Yu, X.; Wei, X.U.; Lin, X. Networking groups opinion leader identification algorithms based on sentiment analysis. *Comput. Sci.* **2012**, *39*, 34–37.
22. Yin, C.; Liang, C.; Yen, F. Determinants of public attitude towards a social enterprise crisis in the digital era: Lessons learnt from THINX. *Public Relat. Rev.* **2018**, *44*, 784–793.
23. Zhang, L.; Zhang, N. Effectiveness of trust repair strategies in the crisis of corporate internet public opinion. *Am. J. Manag. Sci. Eng.* **2020**, *5*, 10. [CrossRef]
24. Bella, A.L.; Colladon, A.F.; Battistoni, E.; Castellan, S.; Francucci, M. Assessing perceived organizational leadership styles through twitter text mining. *J. Assoc. Inf. Sci. Technol.* **2018**, *69*, 21–31. [CrossRef]
25. Liu, D.; Su, J.; Song, L.; Qiu, Z. Application of internet segmentation research based on natural language processing technology in enterprise public opinion risk monitoring. *J. Phys. Conf. Ser.* **2019**, *1187*, 42007. [CrossRef]
26. Aggarwal, A.; Singh, A. Geo-localized public perception visualization using GLOPP for social media. In Proceedings of the Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 3–5 October 2017; pp. 439–445.

27. Chen, Z.; Hu, Z. UGC user portrait research. *Comput. Syst. Appl.* **2017**, *26*, 24–30.
28. Chen, T.; Peng, L.; Yin, X.; Jing, B.; Yang, J.; Cong, G.; Li, G. A policy category analysis model for tourism promotion in china during the COVID-19 pandemic based on data mining and binary regression. *Risk Manag. Healthc. Policy* **2020**, *13*, 3211–3233. [CrossRef] [PubMed]
29. Breiman, L. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 2017.
30. Everitt, B.S. Classification and regression trees. In *Encyclopedia of Statistics in Behavioral Science*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2005.