applied sciences

MDPI

Article

Social Media Rumor Refuter Feature Analysis and Crowd Identification Based on XGBoost and NLP

Zongmin Li [1], Qi Zhang [1], Yuhong Wang [2] and Shihang Wang [1,*]

[1] Business School, Sichuan University, Chengdu 610065, China; lizongmin@scu.edu.cn (Z.L.); 2019225020006@stu.scu.edu.cn (Q.Z.)

[2] USC-SJTU Institute of Cultural and Creative Industry, Shanghai Jiaotong University, Shanghai 200240, China; honglexi@sjtu.edu.cn

* Correspondence: sw3275@columbia.edu; Tel.: +86-028-8541-5143

Received: 12 May 2020; Accepted: 6 July 2020; Published: 8 July 2020

Featured Application: Results of this work can be applied to anti-rumor microblog recommendation decisions for social media platforms, in order to reduce the impact of rumors by promoting the spread of the truth.

Abstract: One prominent dark side of online information behavior is the spreading of rumors. The feature analysis and crowd identification of social media rumor refuters based on machine learning methods can shed light on the rumor refutation process. This paper analyzed the association between user features and rumor refuting behavior in five main rumor categories: economics, society, disaster, politics, and military. Natural language processing (NLP) techniques are applied to quantify the user's sentiment tendency and recent interests. Then, those results were combined with other personalized features to train an XGBoost classification model, and potential refuters can be identified. Information from 58,807 Sina Weibo users (including their 646,877 microblogs) for the five anti-rumor microblog categories was collected for model training and feature analysis. The results revealed that there were significant differences between rumor stiflers and refuters, as well as between refuters for different categories. Refuters tended to be more active on social media and a large proportion of them gathered in more developed regions. Tweeting history was a vital reference as well, and refuters showed higher interest in topics related with the rumor refuting message. Meanwhile, features such as gender, age, user labels and sentiment tendency also varied between refuters considering categories.

Keywords: rumor refuter; machine learning; nature language processing; XGBoost; feature analysis

1. Introduction

Because of the widespread popularity of social networks and mobile devices, users are able to immediately exchange information and ideas or access news reports through social media feeds such as Twitter, Facebook, or Sina Weibo [1]. However, the dark side of online information behavior should not be neglected. Due to a general lack of control, incorrect, exaggerated, or distorted information can be easily circulated throughout the networks [2]. This kind of information is defined as a rumor [3] as it does not have publicized confirmations nor official refutations. In all the controversial news stories since Twitter's inception, the rumors were found to reach more people and spread deeper and faster than the actual facts [4]. Rumors have been found to affect a country's public opinion [5], lead to economic losses [6], and even cause political consequences [7]. When the sudden crisis broke out, online rumors were even more popular, seriously disrupting social stability. For example, since January 2020, the epidemic of new coronaviruses has spread, and rumors have emerged on the Internet, causing public panic and anger and intensifying social conflicts.

Combating rumors has been a hot research area. A lot of research focuses on the rumor itself—the identification [8], spread [9–11], and influencing factors of the rumors; while deep-rooted human nature are the main factors for the viral spread of the rumor, that is, people tend to read/share tweets that confirm their existing attitudes (selective exposure), regard information that is more consistent with their pre-existing beliefs as more persuasive (confirmation bias), and prefer entertaining and incredible content (desirability bias) [4]. Existing research on the participants of the rumor is mainly aimed at influential individuals [12] in social networks. At the public level, crowd identification of rumors participants is still worth further study.

When people receive a piece of 'news', they may (1) retweet and comment at the same time, or only retweet (spreaders), (2) deny it and spread a corresponding rumor refuting message (refuters), (3) only comment on it or neglect it (stiflers). Individuals' behaviors are closely related with their attitudes [13]. Lewandowsky et al. believed that the same rumor refutation information should be changed for different opinions and angles according to the characteristics and thinking patterns of different groups of people, avoiding sensitive positions such as political positions and world views [14]. Therefore, given a rumor category, analyzing the characteristics of voluntary refuters, and identifying the special group from all rumor participants, make it possible to design targeted rumors refutation strategies based on the characteristics and thinking patterns of refuters. The application value is that the platform can consciously recommend rumor refutation information to them, even adapt the information to suit their personality. It is of great significance for expanding the acceptance of real news and suppressing the spread of rumors. Understanding the content of rumor refutation is a re-learning process, with great subjectivity and group differences. Therefore, netizens featuring analysis and crowd identification are critical to breaking through rumor governance difficulty. Recently, the rapid developments in deep learning and machine learning methods make it possible to extract and process large amounts of unstructured social media data [15–17], so as to identify different crowds and extract group features. This research topic largely remains unexplored. The only prior work was from Wang et al., who predicted social media rumor refuters only in the disaster category [18].

This study intends to reveal the features of netizens who are willing to retweet rumor refutations (refuters) without extra incentives when confronting rumor refuting messages and user features, and propose a rumor refuter crowd identification model. Five main rumor refuting microblog categories are considered that can potentially affect social stability: economics, society, disaster, politics, and military [19]. Similarities and differences of rumor refuters in these five categories are compared.

Natural Language Processing (NLP) and XGBoost are the main tools in this research. NLP is a subfield of computer science, information engineering, and artificial intelligence, and is concerned with programming computers to process and analyze large amounts of human (natural) languages data [20]. Although NLP is already a mature technology, as far as the authors know, the short text similarities and sentiment analysis have not been well-applied in combating rumors, especially associating them with rumor refuting behaviors. Baidu NLP [21] will be applied to quantify the user microblog content's sentiment (recent sentiment tendency) and similarity with original rumor refuting message (recent interests) as a value between 0 and 1, which can also be viewed as a probability. The higher the value, the higher the probability that the sentiment of the microblog is positive or the microblog content is the same as the rumor refuting message.

XGBoost [22] is a relatively new algorithm that has gained popularity due to its accuracy and robustness. XGBoost utilizes boosting, which trains each new instance to emphasize the training instances previously mis-modeled for better classification results. It is a combination of classification and regression trees (CART) [23], but re-defined the objective function with both training loss and complexity of the trees to decrease the chance of overfitting. Thus, XGBoost is a very strong model with high extensibility.

In recent years, XGBoost has been widely applied to practical problem solving [24,25]. Wang et al. have proved XGBoost was found to be the most efficient machine learning method for disaster rumor refuter identification compared with logistic regression, support vector machines, and random forest [18]. Therefore, this paper chooses XGBoost to construct the potential rumor refuters identification model.

The main contributions of this paper can be summarized as:

(1) The focus of social media users (instead of only considering influential individuals) in the rumor refutation process.

(2) Feature analysis of rumor refuters for five different categories of rumors, which can provide guidance on the personalized recommendation for social media users by accelerating the rumor refutation information dissemination.

(3) XGBoost based identification model to identify the rumor refuters and extract significant features of rumor refuters.

The remainder of this paper is organized as follows. Section 2 gives our research motivations. Section 3 shows the methods and results and Section 4 gives the discussion. Section 5 concludes the work and discusses future research applications.

2. Motivations

Research motivation lies in two aspects.

2.1. Decision-Making Support to Rumor Countermeasures

Identifying rumor refuters based on their features is quite valuable such that social media platforms can recommend rumor refuting microblogs or messages to them as this group is more likely to spread the anti-rumor information and accelerate rumor refutation [18,26]. Although there tends to be far fewer people refuting than spreading rumors [4], this ordinary refuter crowd is considerable still. Due to the potential risk of rumors, it is necessary to develop restraining countermeasures. Most identified countermeasures have been focused on blocking the rumors and spreading the truth [26]. Current practice has tended to seek to identify the influential nodes or opinion leaders to refute rumors, but has neglected the significance of the netizens willing to retweet rumor refutations without extra incentives to convince irrational followers. Therefore, from the perspective of accelerating the truth dissemination process, this paper employs feature analysis and voluntary rumor refuter crowd identification under the hypothesis that if these targeted users can be identified and taken advantage of, it is possible to gain new insights into internet rumor countermeasures.

2.2. Adapting User Features into Rumor Control

Many studies have attempted to identify how the unique features of social network users influence social media behavior. It was essential for personalized recommendation systems to detect the accurate and targeted user properties [27]. There were multiple social network identities such as microblog authors, stiflers, and retweeters. For example, some researchers collected potential author attributes such as gender, age, regional origin, and political orientation and found some feature-based differences [28]; others differentiated the features of stiflers and retweeters, and concluded that the stiflers were more concerned about social relationships and the retweeters were more driven by message content [5].

All of this prior research contributed to this paper’s feature set construction. In addition, user retweet histories, status, active time, and interests also impacted retweet behavior. Hence, the similarities between the content of the target tweet and past retweeter posts [29] and users’ subjective feelings were also determining factors [30]. However, few works have been done in adapting those user analyses into rumor control and refuter identification. Previous investigations have involved social media platforms with different user structures (i.e., Twitter and Facebook), but the conclusions could not be generalized to microblog users. Overall, few studies on retweeter attributes have commented on the distinctions between the different original microblog types; therefore, this paper analyzes the refuter features based on anti-rumor classifications.

3. Methods and Results

In this section, we present the methods and relevant results in detail. The overall framework of the methods is shown in Figure 1. Firstly, the data are collected and cleaned. Then, the gender and label frequency comparisons between refuters of five categories are made, which form a rough refuter portrait. Thirdly, sentiment analysis and short text similarity analysis are made. If there are missing values in microblogs/label information/verified information/signature, a value of 0.5 will be assigned. Based on the trained XGBoost classification model, the refuter feature analysis is conducted.

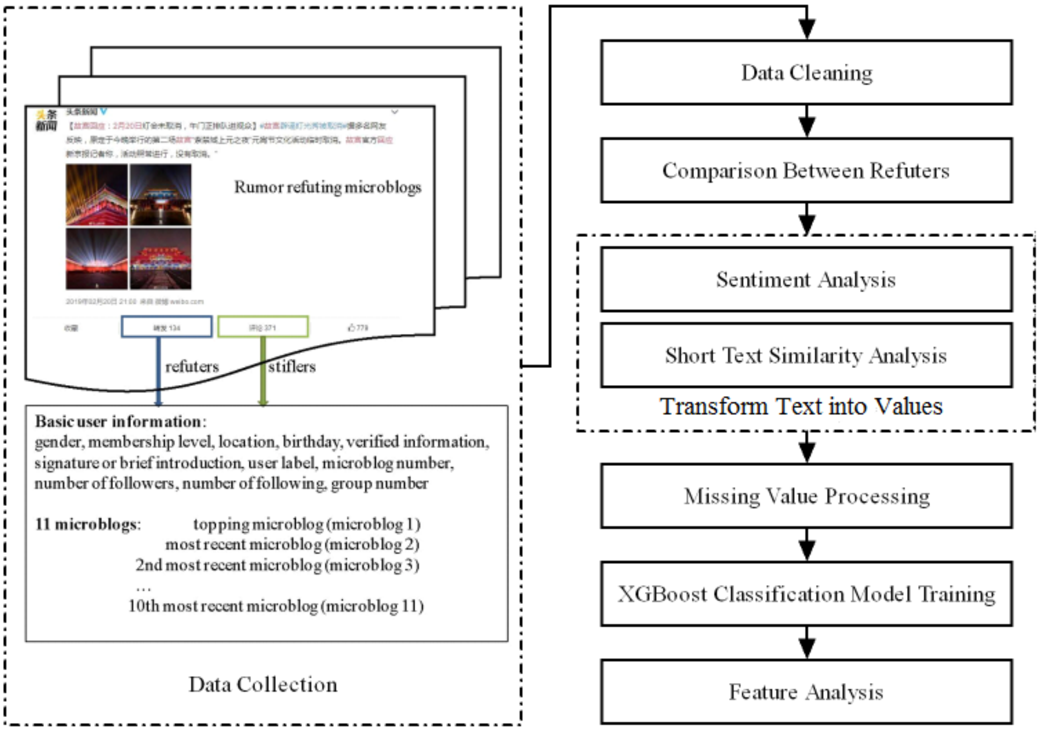
Rumor refuting microblogs
refuters
stiflers
Basic user information:
gender, membership level, location, birthday, verified information, signature or brief introduction, user label, microblog number, number of followers, number of following, group number
11 microblogs: topping microblog (microblog 1)
most recent microblog (microblog 2)
2nd most recent microblog (microblog 3)
...
10th most recent microblog (microblog 11)
Data Collection
Data Cleaning
Comparison Between Refuters
Sentiment Analysis
Short Text Similarity Analysis
Transform Text into Values
Missing Value Processing
XGBoost Classification Model Training
Feature Analysis

Figure 1. The overall framework of the methods.

3.1. Data Collection

Sina Weibo is a Chinese microblogging (Weibo) website and is one of the most popular social media platforms in China with 431 million active monthly users in 2019. Different from Wechat, which only allows a user to post to certified friends, Sina Weibo is the Chinese equivalent of Twitter as it has a wider, more open dispersal. Therefore, crawling microblogs on Sina Weibo has a high research value for rumor propagation or rumor refutation spread analyses. All of the anti-rumor microblogs with a retweet/comment amount larger than 100 were collected from October 2018 to April 2019 using a web crawler.

This paper only takes the anti-rumor microblogs verified, confirmed, and announced by official accounts (police accounts, government agency accounts, and authoritative media accounts) into considerations. Therefore, the refuters discussed in this paper are those who deliberately spread official accounts' rumor refutation information. As shown in Figure 2, the collected anti-rumor microblogs were classified into five categories based on content [18]; economics, society, disaster, politics, the military, all of which were the common rumors on social media platforms and could result in societal damage. The economic category contained business and entrepreneurial information; the society category covered rumors about social public affairs; the disaster category consisted of distorted information on natural and man-made disasters; the politics category comprised false political messages mainly involving certain political figures, groups, or specific policies; and the military category included rumors about national defense or military affairs.

These five main categories had a total of 106 anti-rumor microblogs, of which 45 were related with the society, 31 with economics, three with the military, 20 with politics and seven with disaster, with a total of 58,807 user samples. There were far more stiflers than refuters collected because the task of identifying the refuters from the population was inherently an imbalanced classification problem. As this research was simulating the refuter identification process and examining the validity of XGBoost model, testing on a small data subset was considered powerful enough to examine the algorithm's performance [31].

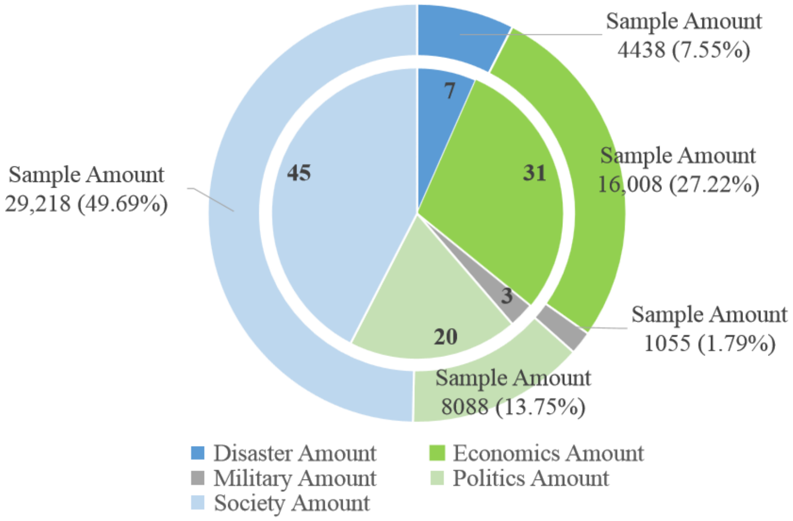

Figure 2. Microblog and sample quantities.

The users' most recent concerns were strongly associated with their most recent microblogs and our previous work found that the 11 most recently posted microblogs (topping microblogs are included) were reliable predictors in disaster rumor refuter prediction [18]. Except for a few users who had less than 11 microblogs since registration, 11 microblogs were extracted, i.e., the topping microblog (the sticky microblog) and 10 most recently posted microblogs. Although the topping microblogs might not have been recently posted, they were able to reveal the overall attitude of the users to some extent. Basic user information; gender, membership level, location, birthday, verified information, signature or brief introduction, user label, microblog number, number of followers and numbers following, and group numbers for each user; were extracted.

As the aim of this research was to identify the social media rumor refuter features, information was mined for two groups of people: refuters from the retweet lists and stiflers from the comment lists. Stiflers consist of the commenters and the users who only view the rumor refuting message. Due to the inaccessibility of the viewer list, only those commenters were treated as stiflers. Although both of refuters and stiflers had viewed the rumor refuting microblogs, the responses were quite different.

3.2. Comparison between Refuters

3.2.1. Gender

Figure 3 compares the gender differences for different categories. The gender gap was particularly large in the military and political fields, with the number of male refuters being nearly twice as many as females (roughly 150 men vs. 74 women and 1621 men vs. 888 women, respectively). In contrast, in the economic, disaster and society categories, there were only minor gender differences. In 2018, male users made up 57% of total Weibo users [32]. The results are correspondent with the Weibo user gender ratio.

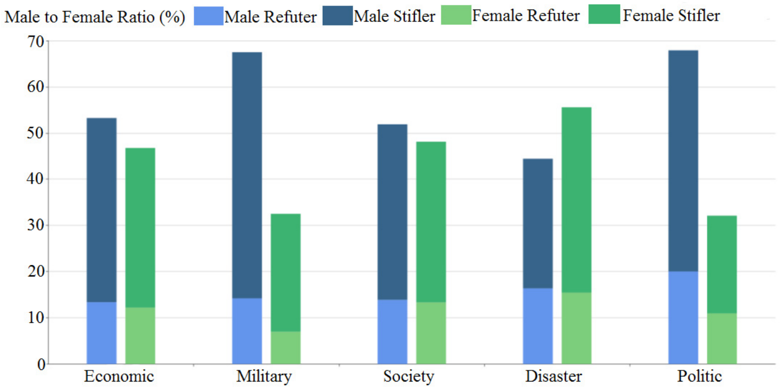

Figure 3. Male to female ratio in database.

3.2.2. Label Frequency

A user portrait analysis was conducted based on the refuter label information. From the word frequency count, it was possible to roughly depict the refuter features and preferences.

As can be seen in Table 1, economics-related rumor refuters showed high interest in IT, Dig, and investment, with most being young practitioners in the internet or finance industries. The economic-related and politics-related rumor refuters had some common interests (i.e., military, investment, Finance and IT, and Dig), and could be the same group of people. For the military-related rumor refuters, the label "military" was third ranked, with interest also being shown in design and history. The society-related and disaster-related rumor refuters were also both interested in education, with the former group having a specific "campus" label and the latter group having a specific "employment" label. Based on this information, we infer that, for these groups, college students should account for a relative large proportion of refuters.

Table 1. Label frequency for the different rumor refuters.

Economics		Military		Society		Disaster		Politics	
Label	**Rank**	**Label**	**Rank**	**Label**	**Rank**	**Label**	**Rank**	**Label**	**Rank**
IT&Dig	5	**Military**	3	IT&Dig	8	IT&Dig	10	IT&Dig	6
Invest	12	Invest, Finance	6	**Reading**	14	Invest	11	Finance	8
Military	16	**News**	11	**Fashion**	19	Finance	15	**News**	9
Finance	17	**Design**	12	**Education**	20	**Education**	18	**Military**	11
Reading	20	**History**	15	**Campus**	21	**Employment**	19	Invest	18

3.3. Rumor Refuters Identification

A crowd identification process was applied in two steps.

Step 1. Convert the textual content into numerical values.

For rumors that are linked to specific geographical locations, we derived the locations from the original rumor texts. Then, comparing the locations where the rumor "took place" with the locations each user from, 1 would be assigned for the location feature if the user was in the same province as the rumor, and 0 otherwise.

Baidu's AipNLP [21], which is regarded as the most advanced Chinese text analysis technique, was applied to convert the textual content into numerical values. Then, the similarities between the user labels, the verified information, the signature, the most recent 11 microblog (including the topping microblog) contents, and the rumor refuting microblogs were transformed into values between 0 and 1. For the sentiment analysis, the emotional inclinations of the user signatures and the most recent 11 microblog contents were also converted into values between 0 and 1. The processed variables are listed in Table 2.

Additional implementations were applied to the variables to ensure the classification results were more valid and reliable:

(1) The corresponding rumor refuting microblogs were deleted if they were one of the 11 most recent microblogs from the user.

(2) A value of 0.5 was assigned to the microblogs/label information/verified information/signature sentiment or short text similarity analysis if the text was missing.

(3) Words irrelevant to the content of the text but that significantly influenced the result of the sentiment analysis, such as "Comment", "Like", and "Collect", were removed.

Table 2. Variables for refuter identification.

Variables	Variable Descriptions
R	Dependent variable. Whether the user retweeted the rumor refuting microblog. 1 for yes and 0 for no.
G	Gender of the user set. 1 for male and 0 for female.
ML	Membership Level. Explains the user devotion and activity to some extent.
L	Location. The provincial level location of the user. 1 if the user was in the same province as the rumor; 0 otherwise.
A	Age. The age of the users. If the value was missing, we interpolated the average age in that category.
LSm	Similarity between the label information and the rumor refuting microblog ranging from 0–1; the larger the value, the more similar the two texts.
VISm	Similarity between the verified information and the rumor refuting microblog ranging from 0–1; the larger the value, the more similar the two texts.
SSe	Sentiment of the Signature or a Brief Introduction of the user ranging from 0–1; the larger the value, the more positive the attitude.
SSm	Similarity between the Signature or Brief Introduction of the user and the rumor refuting microblog ranging from 0–1; the larger the value, the more similar the two texts.
NOM	Number of Microblogs the user has already posted.
NOU	Number of Users the user has followed.
NOF	Number of Followers the user has.
NOG	Number of Groups the user has classified as friends.
MSe (1–11)	Sentiment of the ith microblog of the user ranging from 0–1; the larger the value, the more positive the attitude.
MSm (1–11)	Similarity between the ith microblog of the user and the rumor refuting microblog ranging from 0–1; the larger the value, the more similar the two texts.

Step 2. The XGBoost model was utilized for refuter identification in the different categories.

In this research, the XGBoost model got a bi-classification task. Thus, people who forwarded the anti-rumor microblog was treated as rumor refuters and labeled 1, people who only commented but not retweeted was treated as rumor stiflers and labeled 0.

The samples were randomly divided into two parts, 80% for training the XGBoost model and the other 20% for testing the effect of the trained model. Two criteria; the F1 Score [33] and the AUC [34]; were applied for the classification result evaluation (as shown in Table 3). F1 score is a measure of a test's accuracy. It considers both the precision and the recall so that it is practical in classification tasks [33]. AUC is the probability of ranking the positive sample forward the negative sample whenever a positive and a negative samples are randomly selected. The higher the AUC, the better the classification result is [34].

With learning rate of 0.05, max_depth of 10, subsample value of 0.8, scale_pos_weight corresponding to the proportion of positive and negative samples, and keeping all the other parameters as default, the model is trained by Python Xgboost package (num_boost_round = 300 and early_stopping_rounds = 50).

The efficiency of the XGBoost model can also be impacted by the number of samples. Therefore, for the disaster, economic, political, and societal categories and all samples, the amount of samples (randomly selected each time and not applied to the military category because there were only

1055 samples) was gradually increased to examine the influence of the number of samples on the classification results. During this process, different samples were applied for robustness testing and to determine the relationship between sample quantity and the F1 Score/AUC Score when the XGBoost model was applied, with the overall aim being to determine the number of samples needed to obtain a stable, available F1 Score/AUC Score.

The feature importance was also ranked using the XGBoost model to determine the most important refuter crowd identification features for the different rumor categories. For those most important features, *t*-tests were implemented, to identify which individual feature, for instance, number of microblogs or number of followers, was significantly different between refuters and stiflers.

Because the F1 Score and AUC curves were similar, for better observation, only the F1 Score curve was drawn. As it is shown in Figure 4, starting with 500 randomly selected samples, the F1 Score was observed to gradually increase and then, as sample quantity increased in all categories, it became stable at around 0.75 (except for the military category that had only 1055 samples and F1 and AUC scores of 0.65). Even when all sample types were included, the observations remained the same.

Table 3. AUC and F1 Scores for each category and all samples.

Index	Military	Disaster	Politics	Economics	Society	All
AUC	0.6501	0.7404	0.7350	0.7356	0.7304	0.7356
F1 Score	0.6501	0.7488	0.7470	0.7511	0.7446	0.7490

Therefore, it was concluded that, when plenty of data were provided, the XGBoost model was effective in identifying rumor refuters irrespective of the rumor category differences.

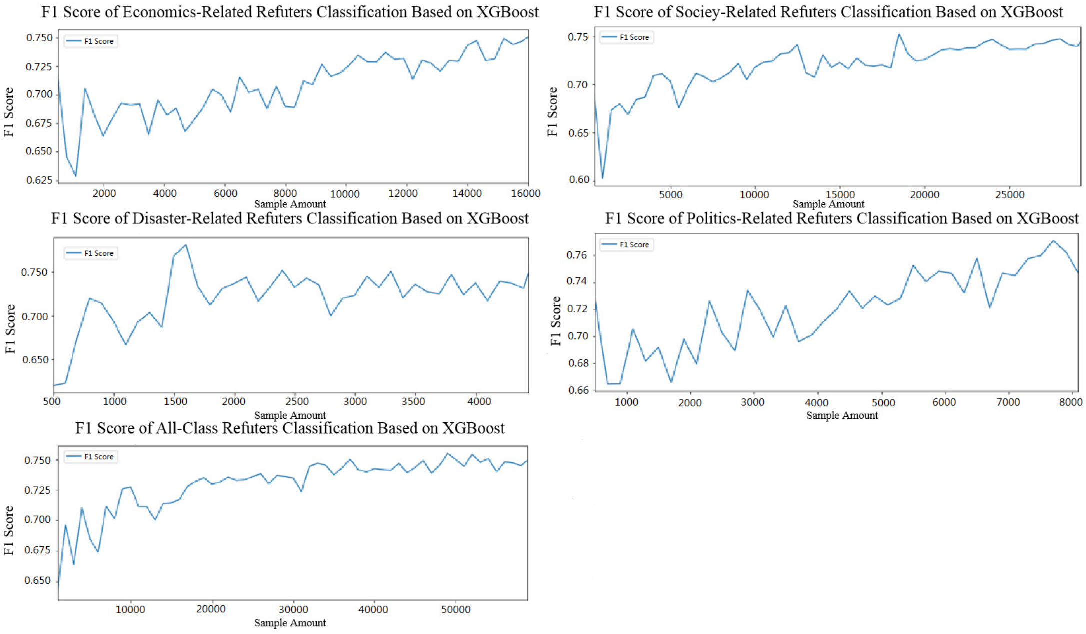

Figure 4. Refuter classification results based on XGBoost.

3.4. Feature Analysis between Refuters and Stiflers

The XGBoost model also provided feature importance rankings (see Figure 5). Except for the disaster and political categories, gender was found to be the least important feature in the XGBoost classifications. However, the MSe11 and MSm11 (regard the topping microblog as the 1st microblog, and MSe11 and MSm11 refer to the sentiment value and similarity with the origin rumor refuting microblog of the 10th most recent microblog respectively) appeared to have the most important features for all categories.

It was therefore proposed that, if the user had the topping microblog and an emotional inclination and similarity to the original microblog, there would be an influence on the classification judgement. Therefore, samples with MSe11 and MSm11 not equal to 0.5 (i.e., samples with topping microblog) were extracted and their Mse1 and MSm1 were tested and the results are shown in Table 4. However, there were no significant differences between the refuter and stifler values for their MSe1 and MSm1 in the political-related, disaster-related and military-related categories. The economics-related and societal-related rumor refuter values for MSm1 were lower than those of the stiflers. There were no significant differences in the MSe1 values between the stiflers and refuters in the economics-related category but, for the societal-related category, the refuters' MSe1 values were higher than those of the stiflers.

The NOM and NOF were also found to be very important features. The *t*-test results in Table 4 found that the rumor refuter NOM and NOF values were somewhat higher than those for the stiflers, which indicated that refuters could be more active. The ML (another measurement of user activity) was also somewhat higher for the refuters in the economics, politics, and societal categories.

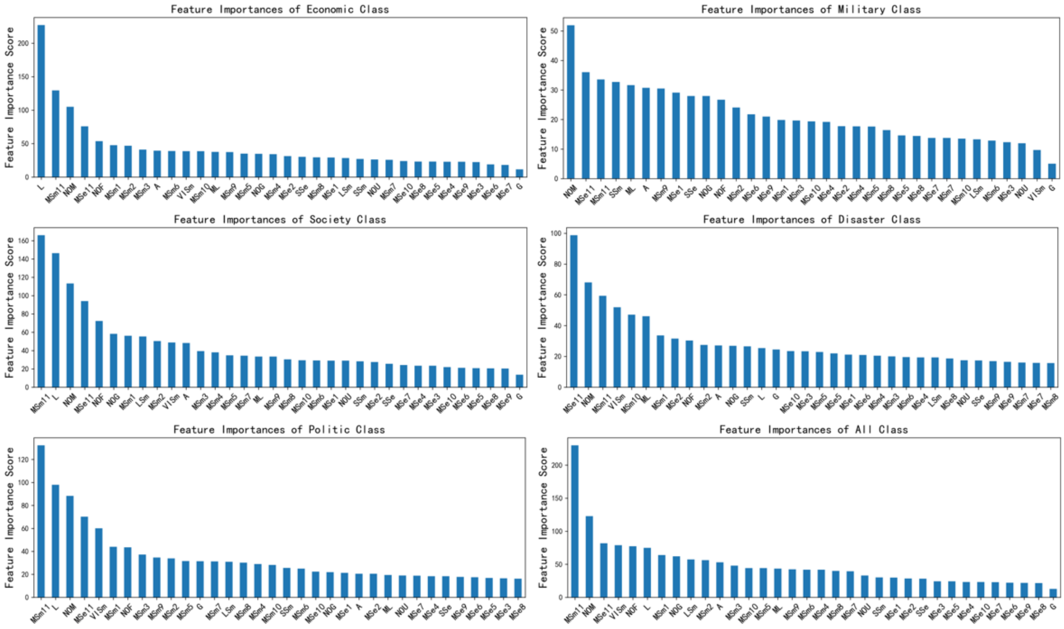

Figure 5. Feature importance in the different categories.

Table 4. *T*-test results for the rumor refuters and stiflers.

	Economic	Military	Societal	Disaster	Political
ML	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
NOM	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
NOF	**0.000**	**0.003**	**0.000**	**0.000**	**0.000**
VISm	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**
LSm	**0.000**	**0.033**	0.335	**0.000**	**0.000**
SSm	**0.000**	**0.000**	**0.000**	**0.027**	0.815
MSe1	0.501	0.526	**0.001**	0.792	0.215
MSm1	**0.023**	0.738	**0.001**	0.540	0.439
$\overline{\text{MSe}}$	**0.001**	0.526	0.512	0.818	0.240
$\overline{\text{MSm}}$	**0.000**	**0.003**	**0.000**	**0.000**	**0.000**

Note: (the value is shown in bold if there were significant differences between the refuters and stiflers under a 95% confidence level, "___" means that the refuter value was lower than that of the stiflers; for users with a topping microblog, the MSe1 and MSm1 refuter and stifler values were compared.).

Except for the LSm in the societal category and the SSm in the political category, there were significant differences found for the VISm, LSm, and SSm between the refuters and stiflers at a 95% confidence level. For the economic, disaster, and military categories, the VISm and LSm of the refuters were significantly lower than those of the stiflers, while the refuters had higher SSm values. Similarly, for the societal category, the VISm of the refuters was significantly lower than that of stiflers, while the refuters had higher SSm values. In contrast, in the political category, the VISm and LSm of refuters were significantly higher than those of the stiflers.

The average values for the MSe (1–11) and the MSm (1–11) for the refuters and the stiflers in the 5 main categories were calculated and denoted $\overline{\text{MSe}}$ and $\overline{\text{MSm}}$. As shown in Table 4, there were no significant differences found between the refuters and the stiflers for the $\overline{\text{MSe}}$ at a 95% confidence level, except for the economic category (the $\overline{\text{MSe}}$ of refuters was significantly lower than that of stiflers). The $\overline{\text{MSm}}$ of the refuters in all five categories, however, was higher than that of the stiflers, which indicated that the average short text refuter similarity degrees with the original rumor refuting microblogs were significantly higher than those of the stiflers.

According to Table 5, at the 95% confidence level, in the disaster, economic and society related rumor refuting microblogs, correlations were confirmed between user behavior (refute/stifle) and user location (whether in the same province in which the rumor-related event occurred); however, in the political category, no correlations were found.

Table 5. Chi-square test of contingency results between user behavior and user location.

Category	Degree of Freedom	Chi-Value	*p*-Value
Economic	1	12.6312	0.0004
Societal	1	13.6454	0.0002
Disaster	1	65.7010	0
Politic	1	2.8048	0.094
Category & Behavior	4	142.3592	0

As shown in Table 6, for users in the same location as the rumor refuting microblogs, the refuters were found to be less likely to retweet disaster, economic or society related rumor refutation information, with only 21.56%, 15.67%, and 23.77% of total viewers in the same province. One possible explanation is that these refuters know better about the local situation and do not feel the urge to spread truths. Therefore, the social media platform can recommend disaster, economic, or society related anti-rumor information to users not in the same location as the rumor refuting microblogs.

Table 6. Refuter proportions in the same province as the anti-rumor microblogs.

Category	Disaster	Economic	Military	Politics	Society
Refuters in the same province	224	39	0	42	498
Stiflers in the same province	815	210	0	66	1597
Refuter proportion in the same province	21.56%	15.67%	-	38.89%	23.77%
Refuters not in the same province	1190	4062	224	2467	7466
Stiflers not in the same province	2209	11,697	831	5513	19,657
Refuter proportion not in the same province	35.01%	25.77%	21.23%	30.91%	27.53%

4. Discussion

Based on feature analysis of users with different social media behavior, this study sought to identify the potential voluntary rumor refuter, and utilize them with the anti-rumor countermeasure: truth propagation and targeted immunization. Because of the growing popularity of social media and the availability of complete user information, it is possible to accurately obtain user features and therefore easier to identify the potential refuters. Thus, personalized recommendation services could be provided to trigger the spread of the truth, and thus enhance rumor refutation.

Although previous works have explored the features of retweeters, there have been few studies on utilizing these findings to combat rumor spread. This paper extended the scope of current studies, instead of studying the general features of retweeters or the opposite group, rumor spreaders, it focused on refuters and specified them with five main rumor categories that can affect social stability. Although both rumor spreaders and rumor refuters have the same behavior—retweet, their features were different. In contrast with the conclusion of Vosoughi et al. [4], in which the rumor spreaders were found to have less followers, it was observed that the rumor refuters had a greater number of microblogs and followers; i.e., they were more active. However, this result could be partially explained by Zhang et al. [35] that social relationship and message content were noticeable driven factors of retweeting behavior. Our findings were also in line with the literature indicating that users mainly retweet to remind others and express themselves, and retweetability is closely related to the number of followers and tweet contents' information and value [36]. It can be recognized that, when user got more followers, they tended to be more cautious with their microblog contents. Thus, retweeted messages that seem more reliable, and rumor refuting messages released by authoritative media could be one of those.

Except for the economic category, there were no significant general sentiment tendency differences between stiflers and refuters. However, the microblog contents and signature contents (except for the political type) of refuters got higher similarity with the original rumor refuting message, and this result was consistent with Luo et al. [29] and Macskassy and Michelson [37]. On the contrary, the similarity between rumor refuting message and verified information and user label were generally lower for refuters (except for the political type), which indicated that the circles and occupations of users were not seriously constant with their daily interests on the social media. Meanwhile, refuters tended to gather in more developed regions.

There were specific rumor refuter feature variations in the microblog categories that had not been previously detected. The politic and military related rumor refuters were generally older and many of them showed interests in finance and investment. Oppositely, the younger ones were more likely to be economic, society, and disaster related rumor refuters. Many of them showed interests in IT&Dig, reading, fashion, education, and employment, and those labels matched their age well. Users in the same province with the rumor seemed less likely to retweet the rumor refuting message. This phenomenon could be explained by the third person effects [38]. On the one hand, the more negative the event was, the more obvious the third person effects were. Due to peoples' underlying sense of superiority and confidence, they unconsciously believe that negative content would exert greater impacts on others than themselves and thus lead to their retweet behaviors to convey information to others (it was also why this phenomenon was most obvious in disaster related rumors). On the other hand, the effectiveness of third person effect is strongly influenced by the geographical distance between the receiver and information source, implying that the farther the receiver is from the information source, the stronger the third-person effect. Therefore, people in other locations thought that retweeting right message was urgent and important, considering those people with both long social and geographical distance would be significantly influenced by media content.

However, the small microblog sample size may have influenced the study's validity to certain extent, and the study was also limited by some of the basic variables that were extracted to characterize the refuter profiles. As the issue of user features has always been intriguing and could be explored from various dimensions, it is expected that, in the future, a wider range of features will be identified in future works to more comprehensively model rumor refuters such as ethnicity, personal preferences, active time, and sociolinguistic features. More empirical studies are also needed to investigate the usefulness and feasibility of the method developed in this paper on other social media platforms such as Wechat, Facebook, and Twitter so that it can be incorporated into active applications.

An additional uncontrolled factor is a difficulty in accurately identifying rumors/anti-rumor on Weibo. In terms of the Chinese legal framework, rumors are generally fake news. This definition emphasizes the deviation from the truth and the fact. From the perspective of mass communication, a rumor is the statement or piece of news that is deliberately made up out of thin air. The malicious motives behind the information source might also be considered. In addition to the rumor itself, there are other forms of information filled up with Weibo, such as uncertain information and speculative information. It can be seen that there is no unified definition of rumors from different academic perspectives, and there are no clear judgment criteria for rumors, so, in practice, many difficulties and problems are unavoidable in the identification of rumors.

The principal purpose of countering the rumors is to filter the literal meaning of rumors, dig into once-hidden problems behind the rumors, and solve the underlying deep-seated social problems reflected, effectively responding to the social anxiety. Given that the main body in China to deal with rumors, solve social problems, and take targeted actions is mostly government agencies, this paper takes whether the false information/refutes of rumors posted by mainstream official accounts (such as police department accounts, government accounts and authoritative media accounts) as the criteria for recognizing and identifying rumors/anti-rumor, so as to maximize the distinction between truth and rumor. Such criteria might still lead to bias in rumors/anti-rumor judgments. Further research might add more dimensions and standards to search and identify rumors/anti-rumor on Weibo, for instance, taking the scientificity, social influence and the poster's subjective intention and other aspects of the web message into the comprehensive consideration.

5. Conclusions and Future Work

The purpose of the current study was to determine the association between user features (including sentiment tendency, recent interests, gender, geographic distributions, age structure, and label frequency) and their refuting behaviors and so as to identify the rumor refuters, and deal with the dark side of online information behavior by accelerating the rumor refutation information dissemination.

The findings shown in Table 7 reveal some general features of refuters as well as variations between refuters considering different rumor categories: (1) there were more male refuters than females, especially in the politics and military categories; (2) rumor refuters of all categories were found to be highly concentrated in East, North and South of China, and particularly in provinces with first-tier cities; (3) when users were from the same geographic locations as the refutation microblogs, they were less inclined to retweet economic, societal and disaster related rumor refutation microblogs; (4) refuters were mainly aged between 18 and 40, with the refuters in the politics and military categories being somewhat older than those in the economic, society, and disaster-related categories; and (5) the political and society related rumor refuters tended to follow and post relevant information more frequently, which was shown by their higher $\overline{\text{MSm}}$.

On the other hand, as it is shown in Table 8, there were significant differences between refuters and stiflers: (1) rumor refuters were found to be more active with higher NOM and NOF; (2) the ML was comparatively higher in the economic, political and societal categories; (3) in general, the refuters' VISm and LSm were significantly lower than the stiflers (except for the LSm in the societal category), but their SSm was higher; however, the refuters in the political category were found to have higher VISm and LSm than the stiflers and there were no significant differences between the SSm of rumor refuters and stiflers; (4) economic related rumor refuters had less positive microblog content sentiment, but the refuters tended to have higher $\overline{\text{MSm}}$ in all categories.

Provided that there was an adequately large amount of data, the XGBoost model was broadly applicable in identifying the refuters, regardless of differences in the rumor categories.

This paper only takes the anti-rumor posts verified, confirmed, and announced by official accounts into consideration, but there is still a small chance that a rumor refuted by the platform turn out to be true eventually. In the future, we plan to examine the refuter characteristics in a wider range of microblog samples, hopefully covering the possible bias with large data. In addition, we will consider more personalized and individualized features beyond just demographic attributes to more precisely identify the refuter crowd. Analysis on influence and power of refuters is also a focus of future research.

Table 7. General features of the rumor refuters.

Category	Common Feature	Gender	Age	Special Interests	Role	Microblog Content	
						Sentiment	Similarity
Economic	Aged: 18–40 ML: 0.89 (mean) NOM: 1363 (median) NOF: 169 (median) NOU: 303(median) Regions: East, North & South China; province with 1st-tier cities (particularly Beijing and Shanghai)	balanced	18.15–37.36	IT & Dig, Invest, Military, Finance	younger practitioners in the internet & finance industries	mean: 0.588 (higher than politic & society)	mean: 0.656
Military		more male	20.14–43.94	Military, Design, History, News	older practitioners in the news industries	mean: 0.575	mean: 0.703
Society		balanced	17.99–39.79	Education, Campus, Reading, Fashion	college students or fresh graduates	mean: 0.558	mean: 0.667 (higher than economic and disaster)
Disaster		balanced	18.08–40.68	Education, Employment	college students or fresh graduates	mean: 0.569 (higher than politic & society)	mean: 0.666 (higher than economic)
Politic		more male	20.32–43.06	IT & Dig, News, Military, Finance	older practitioners in the internet & finance industries	mean: 0.557	mean: 0.674 (higher than economic, society, disaster)

Table 8. Features of the rumor refuters compared with the rumor stifler.

Category	Common Feature	Same Province	ML	Custom Info			Microblog Content
				VISM	LSM	SSM	Sentiment
Economic	Higher: NOM, NOF, $\overline{\text{MSm}}$	less likely to refute	0–2.65 (higher)	0.41–0.55 (lower)	0.14–0.54 (lower)	0.28–0.60 (higher)	relatively less positive
Military		no significant difference	0–1.76 (lower)	0.43–0.55 (lower)	0.12–0.54 (lower)	0.32–0.60 (higher)	no significant difference
Society		less likely to refute	0–2.59 (higher)	0.41–0.55 (lower)	no significant difference	0.28–0.60 (higher)	no significant difference
Disaster		less likely to refute	0–2.55 (lower)	0.41–0.55 (lower)	0.18–0.56 (lower)	0.33–0.63 (higher)	no significant difference
Politic		no significant difference	0–2.76 (higher)	0.41–0.55 (lower)	0.15–0.55 (higher)	no significant difference	no significant difference

Author Contributions: Conceptualization, S.W., Z.L., and Y.W.; Data Curation, S.W., Q.Z., and Y.W.; Formal Analysis, S.W., Z.L., Y.W., and Q.Z.; Investigation, S.W. and Y.W.; Methodology, S.W., Z.L., Q.Z., and Y.W.; Supervision, Z.L.; Validation, S.W., Z.L., Y.W., and Q.Z.; Writing—Original Draft, S.W., Z.L., Y.W., and Q.Z.; Funding acquisition, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the China Postdoctoral Science Foundation Funded Project (Grant No. 2017M612983), Chengdu Philosophy and Social Science Planning Project (Grant No. 2019L40), and the Fundamental Research Funds for the Central Universities (Grant No. SCUBS-PY-202017).

Acknowledgments: We thank the editors and any reviewers for their helpful comments.

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

References

1. Zhao, C.; Xin, Y.; Li, X.; Yang, Y.; Chen, Y. A Heterogeneous Ensemble Learning Framework for Spam Detection in Social Networks with Imbalanced Data. *Appl. Sci.* **2020**, *10*, 936. [CrossRef]
2. Mihailidis, P.; Viotty, S. Spreadable Spectacle in Digital Culture: Civic Expression, Fake News, and the Role of Media Literacies in "Post-Fact" Society. *Am. Behav. Sci.* **2017**, *61*, 441–454. [CrossRef]
3. Difonzo, N.; Bordia, P.; Rosnow, R.L. Reining in rumors. *Organ. Dyn.* **1994**, *23*, 47–62. [CrossRef]
4. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef] [PubMed]
5. Ramos, M.; Shao, J.; Reis, S.D.S.; Anteneodo, C.; Andrade, J.S.; Havlin, S.; Makse, H.A. How does public opinion become extreme. *Sci. Rep.* **2015**, *5*, 10032. [CrossRef]
6. Syrian Hackers' Break into Associated Press' Twitter Account and 'Break News' that Explosions at White House have Injured Obama-Sending DOW Jones Plunging 100 Points. Available online: goo.gl/NSliQP (accessed on 20 December 2019).
7. Humprecht, E. Where "fake news" flourishes: A comparison across four Western democracies. *Inf. Commun. Soc.* **2019**, *22*, 1973–1988. [CrossRef]
8. Liu, Y.; Jin, X.; Shen, H. Towards early identification of online rumors based on long short-term memory networks. *Inf. Process. Manag.* **2019**, *56*, 1457–1467. [CrossRef]
9. Qian, Z.; Tang, S.; Zhang, X.; Zheng, Z. The independent spreaders involved SIR Rumor model in complex networks. *Phys. A Stat. Mech. Appl.* **2015**, *429*, 95–102. [CrossRef]
10. Xia, L.-L.; Jiang, G.-P.; Song, B.; Song, Y. Rumor spreading model considering hesitating mechanism in complex social networks. *Phys. A Stat. Mech. Appl.* **2015**, *437*, 295–303. [CrossRef]
11. Zhang, Y.; Xu, J. A Rumor Spreading Model considering the Cumulative Effects of Memory. *Discret. Dyn. Nat. Soc.* **2015**, *2015*, 1–11. [CrossRef]
12. Goel, S.; Anderson, A.; Hofman, J.; Watts, D.J. The structural virtuality of online diffusion. *Manag. Sci.* **2015**, *62*, 180–196. [CrossRef]
13. Almodarresi, S.M.A.; Tabatabainasab, S.M.; Garabollagh, H.B.; Mohammadi, F. Does citizenship behavior have a role in changing attitude toward green products. *Int. J. Manag. Sci. Eng. Manag.* **2019**, *14*, 284–292. [CrossRef]
14. Lewandowsky, S.; Ecker, U.K.H.; Seifert, C.M.; Schwarz, N.; Cook, J. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychol. Sci. Public Interest* **2012**, *13*, 106–131. [CrossRef]
15. Gholami, P.; Hafezalkotob, A. Maintenance scheduling using data mining techniques and time series models. *Int. J. Manag. Sci. Eng. Manag.* **2018**, *13*, 100–107. [CrossRef]
16. Iglesias, C.A.; Moreno, A. Sentiment analysis for social media. *Appl. Sci.* **2019**, *9*, 5037. [CrossRef]
17. Alotaibi, S.; Mehmood, R.; Katib, I.; Rana, O.; Albeshri, A. Sehaa: A big data analysis tool for healthcare symptoms and diseases detection using twitter, apache spark, and machine learning. *Appl. Sci.* **2020**, *10*, 1398. [CrossRef]
18. Wang, S.; Li, Z.; Wang, Y.; Zhang, Q. Machine Learning Methods to Predict Social Media Disaster Rumor Refuters. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1452. [CrossRef] [PubMed]
19. Wang, G. Dealing with rumors and their control methods from the perspective of communication. *J. Commun.* **1991**, *1*, 41–56. (In Chinese)
20. The History of Machine Translation in a Nutshell. Available online: http://hutchinsweb.me.uk/Nutshell-2005.pdf (accessed on 22 December 2019).
21. Baidu's Aip NLP. Available online: https://pypi.org/project/baidu-aip/ (accessed on 26 December 2019).
22. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
23. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Tree*; Chapman & Hall: New York, NY, USA, 1984.
24. Zheng, C.Y.; Pestilli, F.; Rokem, A. Deconvolution of High Dimensional Mixtures via Boosting, with Application to Diffusion-Weighted MRI of Human Brain. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Nice, France, 2014; pp. 2699–2707.

25. Luo, H.; Schapire, R.E. A Drifting-Games Analysis for Online Learning and Applications to Boosting. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Nice, France, 2014; pp. 1368–1376.
26. Wen, S.; Jiang, J.; Xiang, Y.; Yu, S.; Zhou, W.; Jia, W. To shut them up or to clarify: Restraining the spread of rumors in online social networks. *IEEE Trans. Parallel Distrib. Syst.* **2014**, *25*, 3306–3316. [CrossRef]
27. Khalili-Damghani, K.; Abdi, F.; Abolmakarem, S. Solving customer insurance coverage recommendation problem using a two-stage clustering-classification model. *Int. J. Manag. Sci. Eng. Manag.* **2019**, *14*, 9–19. [CrossRef]
28. Rao, D.; Yarowsky, D.; Shreevats, A.; Gupta, M. Classifying latent user attributes in twitter. In Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents, Toronto, ON, Canada, 26–30 October 2010; pp. 37–44.
29. Luo, Z.; Osborne, M.; Tang, J.; Wang, T. Who will retweet me? Finding retweeters in twitter. In Proceedings of the 36th international ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 28 July–1 August 2013; pp. 869–872.
30. Sun, J.; Wang, G.; Cheng, X.; Fu, Y. Mining affective text to improve social media item recommendation. *Inf. Process. Manag.* **2015**, *51*, 444–457. [CrossRef]
31. Petrak, J. Fast subsampling performance estimates for classification algorithm selection. In Proceedings of the ECML 2000 Workshop on Meta-learning: Building Automatic Advice Strategies for Model Selection and Method Combination, Barcelona, Spain, 31 May–2 June 2000; pp. 3–14.
32. 2018 Weibo User Development Report. Available online: http://data.weibo.com/report/reportDetail?id=433 (accessed on 19 December 2019).
33. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
34. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2005**, *27*, 861–874. [CrossRef]
35. Zhang, J.; Liu, B.; Tang, J.; Chen, T.; Li, J. Social influence locality for modeling retweeting behaviors. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013; pp. 2761–2767.
36. Suh, B.; Hong, L.; Pirolli, P.; Chi, E.H. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In Proceedings of the IEEE International Conference on Social Computing, Minneapolis, MN, USA, 20–22 August 2010.
37. Macskassy, S.A.; Michelson, M. Why do people retweet? Anti-homophily wins the day. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011.
38. Jang, S.M.; Kim, J.K. Third person effects of fake news: Fake news regulation and media literacy interventions. *Comput. Hum. Behav.* **2018**, *80*, 295–302. [CrossRef]