Machine Users Detection on Sina Weibo Platform

Meigen Huang[1],Lihan Zhou[2] and Yu Wang[3]

[1]Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications,China
[2]Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications,China
[3]Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications,China

Abstract. In recent years, the rapid development of Sina Weibo has made it the representative of many Weibo platforms in China. Sina Weibo has attracted large numbers of users in China because of its fast speed of information dissemination, simple use and many star users. More and more Chinese people get news and share information through Sina Weibo. In addition to the normal users, Sina Weibo also appeared on some machine users, these users are generated in order to create false sentiment, which seriously affected the good order of the Sina Weibo platform. By studying normal users and machine users, this paper extracts eight features, they are the number of followings, the number of followers, the number of Weibos, the number of years using Sina Weibo, Sunshine credit, the number of Weibos you like, the proportion of following others by recommending and the ratio of followings and followers. Naive Bayes classification approach, KNN classification approach and SVC classification approach are used for experiment. The experimental results show that the recall rate of the machine users detection is above 96% and the accuracy rate is above 98%, which validates the validity of the features extracted in this paper.

1 Introduction

This article is based on Sina Weibo platform, so Weibo mentioned in this article all refers to Sina Weibo. Sina Weibo is a website launched by Sina network and providing microblog services. In 2009, the radio and television department vigorously rectified the Weibo website, and Sina launched Sina Weibo products, which won the best development period[1]. After the launch of Sina Weibo in August 2009, the number of users showed an explosive growth, breaking through 500 million in just three years[2]. According to the 2017 Weibo user development report, as of September 2017, Weibo has a total of 376 million active users, which has increased by 27% compared with the same period in 2016, of which the mobile terminal accounted for 92%, daily active users reached 165 million, and increased by 25% from last year[3].

In Weibo, which is a huge number of traffic pools, many users want to be famous and want to take advantage of Weibo. The number of fans of users as a key factor to weigh the influence of a Weibo user is also an important reference factor to attract other users to pay attention to themselves. Many marketing users and celebrities are buying fans. The importance of fans has encouraged fans to trade on Weibo. In order to attract other people's attention, some users make their own high popularity by buying fan battalion. In order to make a profit, the sellers have registered thousands of Weibo accounts with software to pay attention to the buyer's Sina Weibo, so as to improve the buyer's false popularity.

As for machine users, there is no authoritative definition in the relevant literature. This article defines the machine users as: on the Sina Weibo platform, the machine users refer to those who are registered and managed by the individual through the software, to create a false and malicious marketing by the means of trading. The Weibo users registered with the software are the traditional escalation of the zombie powder. Most of the traditional zombie users have simple personal information, no head images, and no personal descriptions. But now machine users on the Weibo platform have perfected their basic information in order to avoid Weibo's own malicious user detection measures and disguise them as normal users. These machine users are managed by the individual through the software, creating a false popularity for others, which leads to the inaccurate calculation of the influence of Weibo users, which has brought trouble to the normal users on the Weibo platform. Therefore, it is important to study and detect these machine users to maintain a good order of Sina Weibo platform and to improve the experience of normal users.

2 Present status of related research

The machine users on the Weibo platform mentioned in this article belong to the small branch of the category of spam users in the social network platform. There are many researches on spam users on the network at home and abroad. In literature [4], after crawling the Twitter user data, the spam user in Twitter is annotated with the

a Corresponding author: 969006768@qq.com

method that the hashtag and the content text do not match in the content sent by the spam user, and the spam user is identified in combination with other user characteristics. The literature [5] mainly studied the Weibo content of spam users. In Twitter's spam user identification process, spam users' documents in other media such as E-mail, SMS, and Web, and documents in Twitter were used to form a cross-media knowledge base model to identify spam users in Twitter. M Hull et al. studied the interaction between users in social networks and completed spam recognition in social networks [6]. The research of spam users by foreign scholars is mostly based on the Twitter platform.

The machine users studied in this paper originate from the Sina Weibo platform. There are still some differences between the two platforms. Therefore, it is necessary to study machine users according to the characteristics of Sina Weibo. Wang Yue and others analyzed the different characteristics between zombies powder and ordinary users through the personal information of the Weibo users, the user's Weibo content, and the user's link relationship, and trained a zombie powder classification system based on the C4.5 decision tree[7]. Wang Shuqi and others used the support vector machine algorithm to classify the normal users and the Navy users on Sina Weibo platform, and completed the Navy recognition model[8]. Xie Zhonghong and others used the logistic regression algorithm to identify the navy of Sina Weibo[9]. Liu Yashang and others compared the differences between normal users and zombie users in personal attributes, behavior attributes, content attributes, and relationship attributes, which provided empirical evidence for the research of zombie powder[10].

3 Feature analysis of machine users and normal users

The author purchased 600 Sina Weibo machine user data samples from the Internet, and used the octopus crawler software to collect 300 machine user data samples and 900 normal user data samples. A total of 1800 data samples were used in this paper. Through the comparison of the machine users and the normal users in the sample, it is found that in order to evade the shielding mechanism of the Weibo platform, these machine users will also fill up their own information and will also send the Weibo normally. Therefore, some features extracted from previous studies, such as whether there is a personal introduction and whether there is a head sculpture, the number of Weibo forwarding is no longer applicable to machine user detection. Based on the research of machine user sample data and previous research, eight features of machine user detection are extracted. They are the number of followings, the number of followers, the number of Weibo, the number of years using Sina Weibo, Sunshine credit, the number of likes, the proportion of following others by recommending and the ratio of followings and followers. The former six features have been adopted by many scholars in previous studies. Since these six features are suitable for machine user detection, this paper also continues to use these six features. This article has added two new features, namely the number of likes, the proportion of following others by recommending. These two features have not been applied in the previous literature on Weibo spam detection. The following describes these features:

3.1 Six traditional features

3.1.1 The number of followings

The number of machine users’followings is more than that of normal users, with a range of around 800~3000. In the vermicelli trading, these machine users will pay attention to the buyers' Weibo, so as to raise the popularity of buyers, so these machine users commonly have a large amount of followings. Normal users only concern about their friends or their interested accounts, so the number of followings is very different from those of the machine users.

3.1.2 The number of followers

Since the normal users will pay attention to the people they know, the average user has more fans than the machine user.

3.1.3 The number of Weibos

Normal users share their feelings by sending Weibo or reproducing topics that they are interested in. After research, machine users are controlled by software. These accounts rarely send Weibo or forward to others' Weibo, so the number of machine users’Weibos is less than that of normal users.

3.1.4 The number of years using Sina Weibo

In recent years, the Sina Weibo platform has screened those early machine users through malicious user filtering. So now most of the active machine accounts on Weibo are registered in the last three years. There is a certain difference in the number of years using Sina Weibo between normal users and machine users.

3.1.5 Sunshine credit

Sunshine credit comprehensively considers the credit of Weibo users through multiple dimensions. The level of Sunshine Credit is divided into extremely low credit, low credit, general credit, good credit and excellent credit. It becomes a measure of Weibo users' discussion, positive expression and rational communication on the Internet.

3.1.6 The ratio of followings and followers

Machine users have more followings and less followers than normal users, so the ratio of followings and followers is different.

3.2 Two newly added features in this article

3.2.1 The number of Weibos you like

The number of Weibos you like refers to the number of Weibos which the users like. When browsing other people's Weibo, users will praise their favorite Weibo. Machine users usually don't browse other people's Weibo. Figure 1 below shows the cumulative distribution of the number of likes for normal users and machine users.

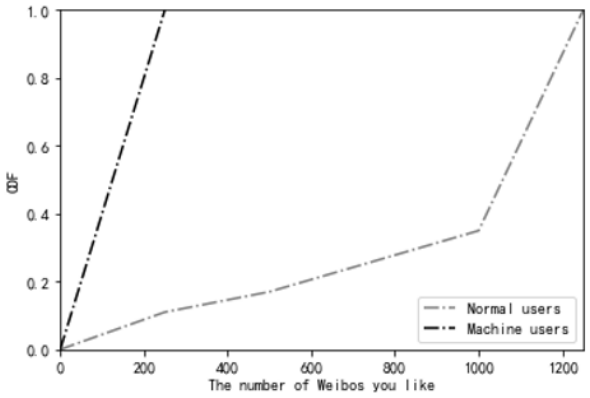

Figure 1. Cumulative distribution of the number of Weibos normal users and machine users like.

As can be seen from figure1, the number of Weibos liked by machine users is below 200. The number of Weibos liked by normal users in the range of 0~1000 occupied 30%, and 70% normal users like more than 1000 Weibos, which indicates that it is appropriate to select the number of Weibos you like as a feature.

3.2.2 The proportion of following others by recommending

There are many ways for Weibo to pay attention to others. You can follow others through finding people or Weibo's recommendation. Generally, normal users will search for the users they are interested in, and they will also follow others through the Weibo's recommendation. After researching the experimental sample data of this article, it is found that the machine users usually follow others through the Weibo recommendation. This is because the machine account is controlled by the software. The fan seller first collects the machine accounts through the group, and then shares the Weibo account of the fan buyer into the group in the recommended way, and finally controls the machine account and then pays attention. This saves time and effort. Due to Weibo's privacy protection mechanism, it is impossible to view all the followings of others. Therefore, this article selects the proportion of following others by recommending in the homepage of normal users and machine users as the feature. Figure 2 below shows the cumulative distribution of the proportion of following others by recommending for normal users and machine users.

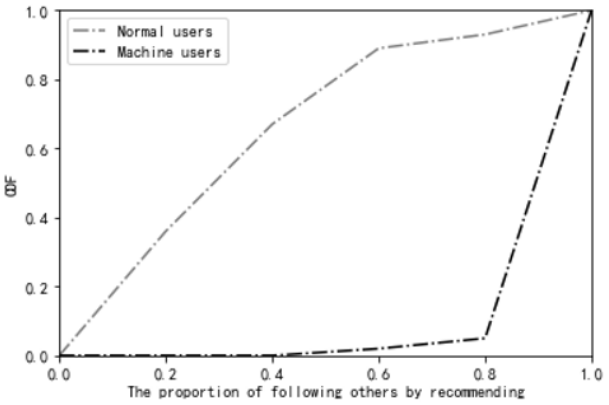

Figure 2. Cumulative distribution of the proportion of following others by recommending for normal users and machine users.

As can be seen from Figure 2 above, the proportion of following others by recommending in the homepage of normal users has risen steadily, with 90% of normal users' proportion being less than 0.6, and 10% of normal users' proportion is between 0.6 and 1. Since the Weibo platform itself will recommend some accounts to users, this small group of users may follow those accounts accidentally. In contrast, 95% of machine users' proportion is more than 0.8. Therefore, this article adds the proportion of following others by recommending to the feature set of the machine user detection.

4 Experiment and analysis

The author purchased 600 Sina Weibo machine user data samples online and used the crawler software to collect 300 machine user data samples and 900 normal user data samples. A total of 1800 data samples were used. In order to verify the validity of the features extracted in this paper, a copy of the data set is duplicated and the number of likes and the proportion of following others by recommending are removed from it.

Based on the Spyder software, three kinds of classification methods, including KNN, naive Bayes and SVC in Scikit-Learn machine learning library, are selected for experiments. In this paper, the data of 1500 users is used as the training data set, and the data of 150 machine users and the data of 150 normal users are randomly selected as the test data set. In this paper, the accuracy rate, recall rate and harmonic mean F1 are used as the evaluation indicators of the classification results.

The results obtained by experiments using the traditional feature set are shown in Table 1 below. The results obtained by experiments using the feature set after the newly added feature are shown in Table 2 below. According to Table 2, we can see that after adding new features, the accuracy of machine user detection using KNN, Naive Bayes and SVC classification algorithms are 99.3%, 98.3% and 99%, respectively. Recall rate, accuracy rate and F_1 value of machine user detection using new feature set are all improved.

Table 1. Experimental results obtained by using traditional feature set.

Classification algorithm	Accuracy rate	Recall rate	F_1
KNN	100.0%	98.0%	98.0%
Naive Bayes	96.6%	96.0%	96.3%
SVC	100.0%	95.3%	97.6%

Table 2. Experimental results obtained by using feature set after adding new features.

Classification algorithm	Accuracy rate	Recall rate	F_1
KNN	100.0%	98.6%	99.3%
Naïve Bayes	100.0%	96.7%	98.3%
SVC	100.0%	98.0%	99.0%

5 Summary

This article carefully studies the data of machine users and normal users on the Sina Weibo platform and extracts eight features, they are the number of followings, the number of followers, the number of Weibos, the number of years using Sina Weibo, Sunshine credit, the number of Weibos you like, the proportion of following others by recommending and the ratio of followings and followers. KNN, Naive Bayes and SVC three kinds of classification algorithms are used for machine users detection, the accuracy of detection is more than 98%, which verifies that the features extracted in this paper are effective for machine users detection on Sina Weibo platform.

References

1. J. Hu, JNR, 48-48(2017)
2. Sh. Zhou, NMR, 8-9(2017)
3. 2017 Weibo user development report, http://data.weibo.com/report/reportDetail?id=404
4. S. Yardi, D.M. Romero, G. Schoenebeck, D. Boyd, (2009)
5. X. Hu, J. Tang, H. Liu, Proceedings of the 37th international ACM,547-556(2014)
6. M. Hull, F. Farmer, E. Perelman, US, US 20050171954 A1 ,(2005)
7. Y. Wang, J. Zhang, F. Liu, CS, 81-86(2014)
8. Sh. Wang, W. Wang, MC,(2018)
9. Zh. Xie, Y. Zhang, L. Zhang, ITNS, 67-69(2017)
10. Y. Liu, B. Chen, H. Zhu, L. Yu, IR, 1-9(2015)