PTT Trolls Terminator

Tsao-Yuan Chang
Computer Science Department
Texas A&M University
College Station, TX
joeychang0204@tamu.edu

Lawrence Lin
Computer Science Department
Texas A&M University
College Station, TX
linlaw0229@tamu.edu

Kuo-Sung Huang
Computer Science Department
Texas A&M University
College Station, TX
s4300713@tamu.edu

ABSTRACT

Online trolls had become a large issue recently. Many discuss forum had been deeply affected by organized trolls or cyber army. In this project, we try to find out such trolls on PTT, which is the largest online discuss forum in Taiwan. We attempt to achieve our goal by combining three methods: IP behavior based threshold filter, comment statistics based threshold filter and content based support vector machine(SVM) classifier.

To build the ground truth, we manually classified the users that made more than fifty comments during a specific time period near the mayoral election. The total number of active user IDs in our dataset is about 3000. Based on the standard we set for trolls, we can reach overall 0.7 F1-score, 0.75 recall, 0.66 precision and 0.79 accuracy.

KEYWORDS

content-based SVM, IP analysis, trolls, supervised learning

1 Introduction

Politically motivated cyber-attacks have increased in recent years. But in almost all cases, they've generally been attributed to unidentified political hackers. PTT Bulletin Board System is the most popular online forum in Taiwan, and it has always been an open-sourced platform which allows people to share thoughts across the country. As the increased number of self-media, evidence shows that more and more political parties are hiring trolls (or 網軍, cyber army) to write scandals and rumor posts in order to attack their opponents. Considering the increasing influence of the social media, many political parties try to do spin control on forums. In fact, there're also strong evidence showing the intervention of offshore funds from China. The troll issue becomes extremely annoying when elections are approaching, so we'd like to integrate a content-based model with other features to classify the trolls on PTT. In this work, we tried to investigate the topic via comment statistics analysis, IP statistics and analysis, and content-based SVM model. The result shows that we can successfully identify the trolls based on some features. However, there are still plenty of things to be explored on this topic.

2 Related Works

Since the trolls are causing a terrible user experience, there were some previous works which also tried to deal with this problem. For example, [1] is a tool which can help us find the users who are sharing their IP address. Given an user's ID, it'll show all the other users who had shared their IP address with this ID and how many times they had shared. This is an interesting project which makes it a lot easier to target some malicious users and find their sockpuppet. However, it's a time-consuming process to identify trolls using this method. We'll have to first find a user acting strange and then analyze the behavior and correlation of this user and his neighbors, which is usually done manually. In our work, we try to find the trolls in a different way. Most of the time, we predict whether a user is a troll based on the comment content, while we also use the comment length and number of used IP as a support. By using our method, we can find the trolls automatically if we can capture the trolls' behavior successfully.

To retrieve data from PTT, we made use of [2], a crawler which can crawl the articles and comments from a specific board on PTT. [3] is a helpful reference to perform Chinese Word2Vec. We use a similar approach as [3] to

convert comments into vectors : first we use [4] to split the comments into words, and then we apply this Chinese Word2Vec to get the vectors. Finally we use our labeled training set and these vectors to train our SVM classifier.

classified as trolls. As a result, 835 (about one-third) of users are classified as trolls while the others are normal users. Among the 3020 popular articles, 2200(about two-third) of them are political articles.

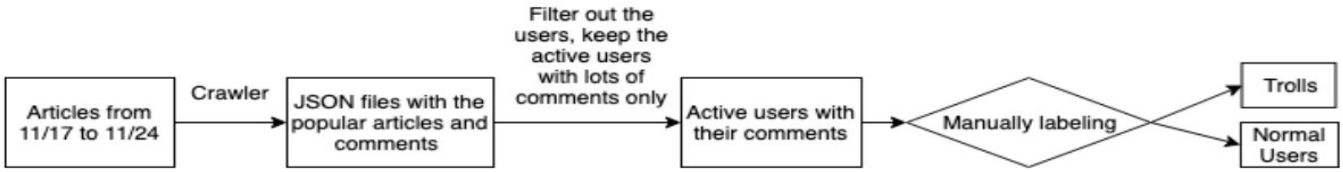

Fig 1. Flow chart of building ground truth

3 Method

In our research, we can divide the work into three different parts. First, how does a given user interact with others? We aim to identify if a user is a troll from the average comment length under articles and the response time he responded to political articles. Second, who is this user and where is he from? We extracted the IP information from the raw data and calculate the statistical numbers of different groups of users. Also, since we're able to find the geolocation of an IP, we summarized the country where a user logged in as well. Third, what kind of thought or content did this user share with others? We make use of the content-based model to classify if a user belongs to trolls. The upcoming sections elaborate our work as follows. First, we explained the process of data collection in 3.1. In 3.2, we explained how we retrieve the comment length and how to set up threshold for response time. In 3.3, IP processing and IP statics analysis is elaborated. In section 3.4, we showed how we develop the content-based SVM model. Then in the last section, we explained how we integrate these three different features together.

3.1 Data Collection

In Figure 1, we showed how we collect our ground truth dataset. Although PTT is an open-source forum, it does not have the troll labels that we need. Therefore, we manually labelled 2747 users. Since we assume the trolls will be more active when election approaches, we first crawled all the articles and comments in the period of one week before the mayoral election in 2018. From this data, we selected the popular articles with a lot of comments, and manually classify whether these articles are political-related. On the other hand, we manually labeled the users who made a lot of comments as trolls or not as our ground truth. Users who with obvious political tendency while attacking others or urging others to vote for somebody would be

3.2 Comment length and response time

In this section, we try to find out how a user interact with others from the comment length and response time. While labeling the users, we discovered that lots of trolls kept making short comments like 'agree' or 'disagree'. If the trolls are hired by someone and they control a lot of accounts, they may want to leave short messages and quickly switch to another account. Also, considering the trolls may try to do spin control as soon as the political articles are posted, we consider the users' average response time as one of the potential feature. Our result shows that users with a very low average comment length are very likely to be trolls while the response time of normal users and trolls shows little difference. When calculating the comment statistics, since we only care about the comments under the political related articles, we built an article classifier. We simply first classify every articles that mentioned the name of politicians as political related. Then manually label the rest of them by human eye.

3.3 IP processing

One of our goals is to find who is the user, where this user comes from, and extract if there is any user sharing the same location with him. In the *Gossiping and HatePolitics* discussion board in PTT, we can view the IP information of each user who posted the article and comments. Hence, the feature is composed of User IP extraction and IP region analysis. The idea of User IP extraction is to find the suspicious users based on their login information. From the raw data we collected, we're able to extract the IP address from each user if they post or commented in the forum. We collect the statistics of the

suspicious users and normal users in the training set. If a user used a relatively large amount of IP addresses, we'll identify it as a troll.

On the other hand, we notice that which region the user is logging from is also important. Since PTT is now considered being compromised by lots of users who support Chinese Unification, we suspect some of the user would log into PTT from China. We applied the python package[5] to retrieve the geographic information of the given IP.

3.4 Content-based SVM model

The content-based model can be divided into three steps. The first step is segmentation. For each of the comments in training set, we use python jieba module with own build special word dictionary to do segmentation. The second part is word to vector. We use the well known word2vec module to transform the segmented comments and articles to vector. Finally we use manually labeled training set data and the transformed data word vector to train a SVM model.

3.5 PTT Trolls Terminator

After some observation to the result of the three method we use to predict the class of the users, we find that comment length and IP process both has very good precision and accuracy yet very low recall. This means that users predicted as trolls by them are very likely to be trolls but we can only find few of these trolls. So we decide that for each id and its comment, we first filter it by IP behavior and comment length. If these methods predict the user as a troll, then we predict it as a troll. Otherwise, if they predict the id as an normal user, we then use SVM model to do the final decision. Fig. 2 shows the flowchart for building PTT trolls terminator.

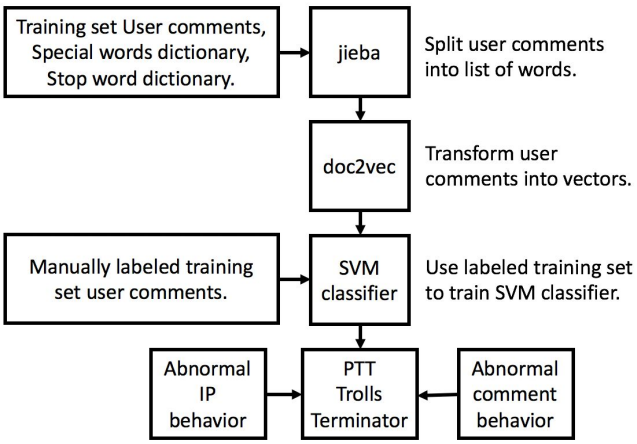

Figure 2. The flowchart of building PTT trolls terminator.

4 Evaluation

To evaluate the performance of PTT trolls terminator, we divide the manually labeled user ids into training set and test set randomly. We used the training set to calculate the statistics like number of used IP, comment length and response time, and train the SVM model. Then we used the test set to evaluate the performance of PTT trolls terminator.Table 1 shows the evaluation result.

Method	Accuracy	Precision	Recall	F1-Score
Comment Length	0.72	0.97	0.17	0.28
Number of IP Used	0.69	0.76	0.10	0.17
SVM	0.77	0.65	0.66	0.66
PTT Troll Terminator	0.79	0.66	0.75	0.70

Table 1. Performance of each method.

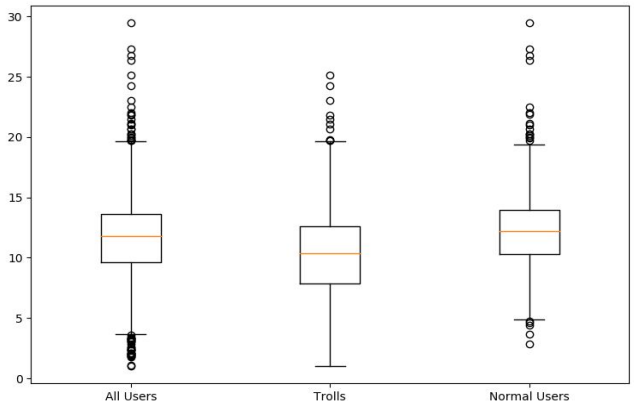

Fig 3.Boxplot of Comment Length

Figure 3 demonstrates the difference in comment length between trolls and normal users. Lots of trolls made extremely short comments such as 'agree' or 'disagree' under political articles, making the comment length distribution for trolls become shorter than normal users. Thus, we take the comment length as one of the useful features and filter out the trolls with extremely short average comment length.

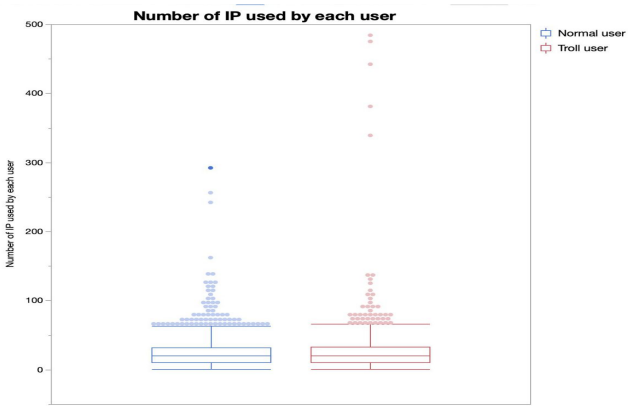

Fig 4.Boxplot of number of IP used per user

Figure 4 shows the boxplot of number of IP used by each user. It shows that troll users(red plots) tends to use more IP than normal users. We think the reason is that the trolls try to hide their internet footprint using VPN.

5 Discussion

From the result, we can see that the filter using number of used IP has 0.69 accuracy and 0.76 precision, which are pretty decent numbers. However, the recall is only 0.1. Comment length filter performs similarly. It has 0.72 accuracy and 0.97 precision yet only 0.17 recall. The reason for this is the threshold we choose is very strict. The users would be classified as trolls only if the number of used IP is higher than average plus two times deviation of normal users and the average length of comments is lower than average minus two times deviation. Combining the three methods, we can achieve an overall 0.7 F1-score, 0.75 recall, 0.66 precision and 0.79 accuracy.

When we tried to figure out what method we should use to identify the trolls, besides the success methods, we also had some failed attempts. The following figures are the failed features.

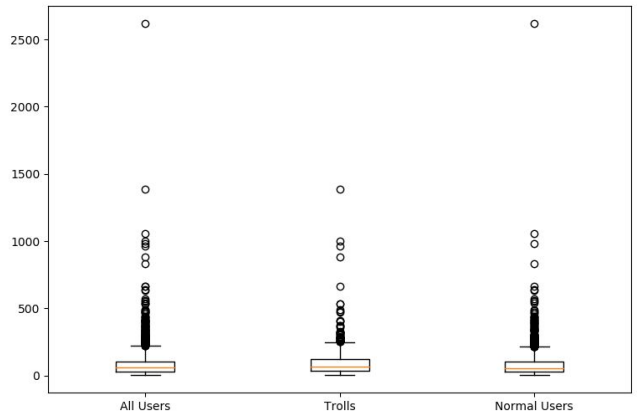

Fig 5.Boxplot of Response Time

One of our intuition was that trolls may want to reply to political articles as soon as possible to do spin control. However our result shows that between trolls and normal users, there's no big difference in response time. This observation suggests that maybe most of the users like to reply to the articles they're interested in as soon as they see the articles, and the response time may not be a good feature to predict trolls.

Fig 6.Region of user login

From Figure 6 and Table 2, we showed that we try to investigate the IP region of users. However, the result is not satisfied. It shows that most of the users are from Taiwan and only a few of them are from other countries.

Top 10 Country from mainly used IP	
TW	89083(95.5%)
US	1216(1.3%)
JP	867(0.93%)
AU	304(0.33%)
HK	236(0.25%)
CN	207(0.22%)
SG	153(0.16%)
VN	147(0.16%)
GB	125(0.13%)
CA	118(0.13%)

Table 2. Top 10 Region of user login

6 Conclusion

Building ground truth is the most difficult problem in our work. The user's intention is difficult to infer even if we use human eyes to check the comments one by one. Based on the standard we set for the trolls as the ground truth, we perform three different methods to find them out. Content based SVM model has over all F1-score 0.66, 0.66 recall and 0.65 precision. IP analysis has only 0.17 F1-score score and 0.1 recall yet 0.76 precision. Similar to IP based filter, comment length filter has 0.28 F1-score and 0.17 recall yet 0.97 precision. Combine the three methods, we can achieve an overall 0.7 F1-score, 0.75 recall, 0.66 precision and 0.79 accuracy. We can find the trolls with a decent possibility but the model still need more improvements.

To further improve and explore this project, we have several ideas which may make the whole system more convincing. Currently, we labeled the users into two classes: trolls or normal users. One potential future work is to make a more detailed classification. For example, we can further separate the trolls according to who they're supporting or who they're attacking. Another work we can do is to extend our project to the latest data. Our initial goal was to distinguish whether a given user is a troll or not, while for now we can only classify the active users in the time period before mayoral election. However, as the data grows, we'll need some better technique for data labeling. Maybe we should design a concrete criteria to make the labeling process more consistent. Finally, an interactive interface may help us demonstrate our work better. A web application which takes user ID as input and outputs the user's statistics and our troll prediction is a direction we can work on in the future.

REFERENCES

[1] Analyze user IP in PTT, https://www.ianalyseur.org/

[2] PTT-web-crawler, https://github.com/jwlin/ptt-web-crawler

[3] Word2vec, Word2vec was created by a team of researchers led by Tomas Mikolov at Google https://github.com/Alex-CHUN-YU/Word2vec

[4] Chinese text segmentation: built to be the best Python Chinese word segmentation module https://github.com/fxsjy/jieba

[5] MaxMind GeoIP2 Python API, https://github.com/maxmind/GeoIP2-python