PTT Trolls Terminator

Tsao-Yuan Chang, Kuo-Sung Huang, Lawrence Lin
Texas A&M University

Introduction

PTT Bulletin Board System is the most popular forum in Taiwan, which acts like Reddit. Thus some political parties are hiring trolls (or 網軍, cyber army) to write some scandal or rumor posts in order to attack their opponents. Since the PTT has founded, it helped the public to post and share their thoughts unrestrainedly. As the influence of the social media increased, many political parties try to do spin control on the forum. In fact, there're strong evidence showing that offshore funds, for example China, also part of the chess. 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. 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.

Methodology

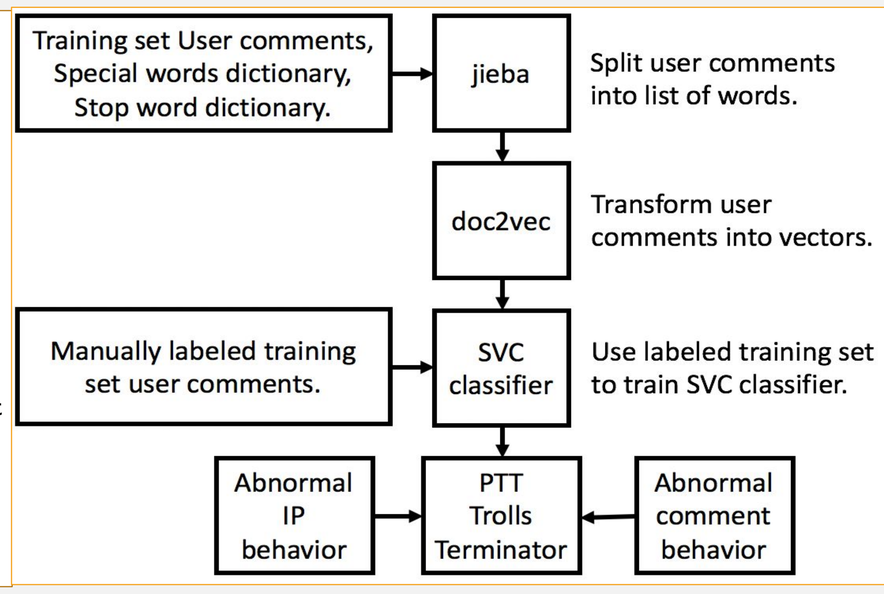
Our approach can be divided into three parts: response time analysis, IP analysis, and content-based model training. The response length and IP analysis helped us to classify if user X is a troll based on the preliminary historical results.

Content-based model training: 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.

Comment length and response time: While labeling the users, we discovered that lots of trolls kept leaving 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 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.

IP processing: This feature is composed of User IP extraction and IP region analysis. The idea of User IP extraction tries 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/commented in the forum. We collect the logistics 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 think the region of where this user login is also important. Since PTT is now considered been compromised by lots of users who support Chinese Unification, we suspect some of the user would logged into PTT from China.

To combine the three methods, we first found that IP model and comment model have very high precision yet low recall. So we decide to first filt the data by these two models. If a user is classified as troll by them, we predict it as a troll. If they predict it as normal user, we then use content-based model to predict.



Data Collection

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. 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 other users would be classified as trolls.

Fig 1.Screenshot of PTT

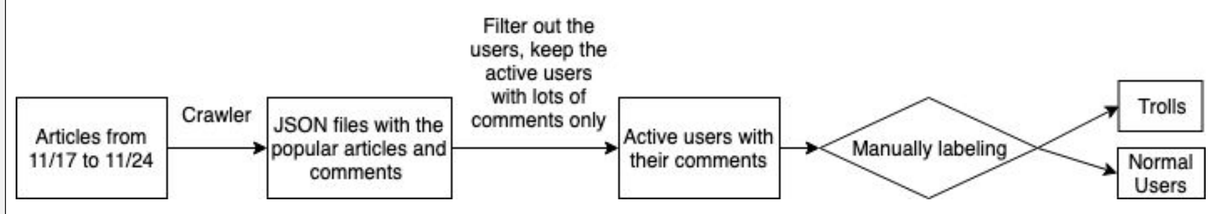


Fig 1. Flow chart of building ground truth

Result

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

Table 1. Performance of each method.

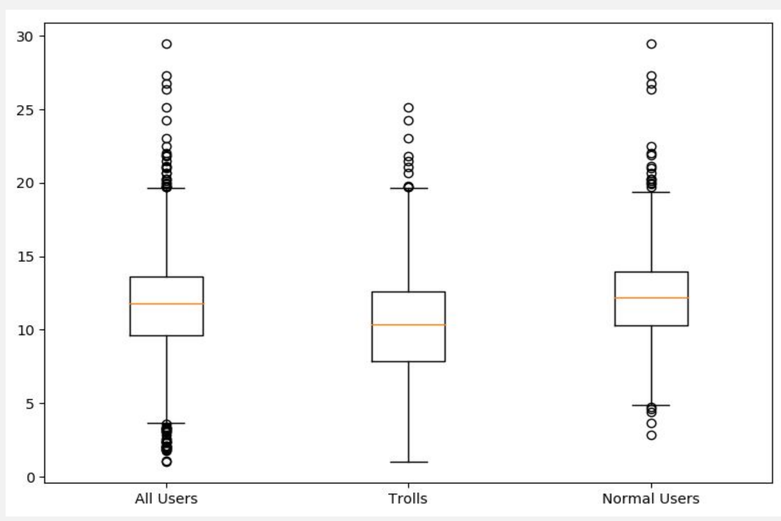


Fig 3.Boxplot of Comment Length

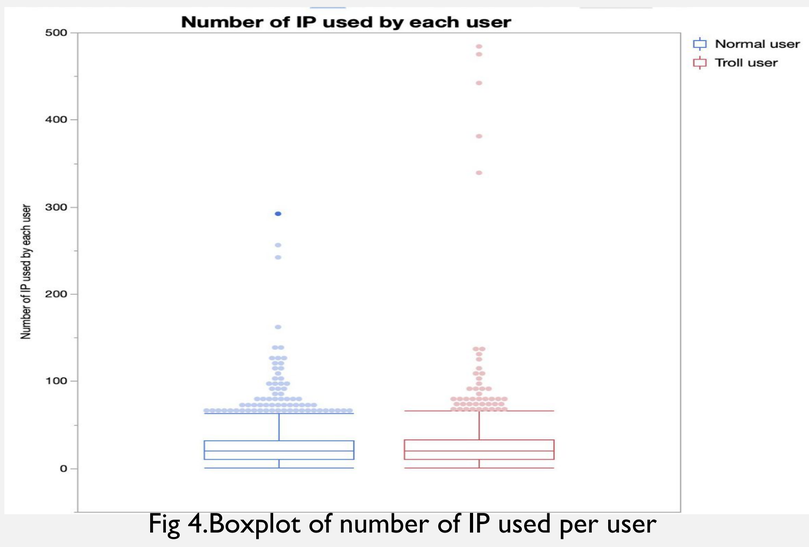


Fig 4.Boxplot of number of IP used per user

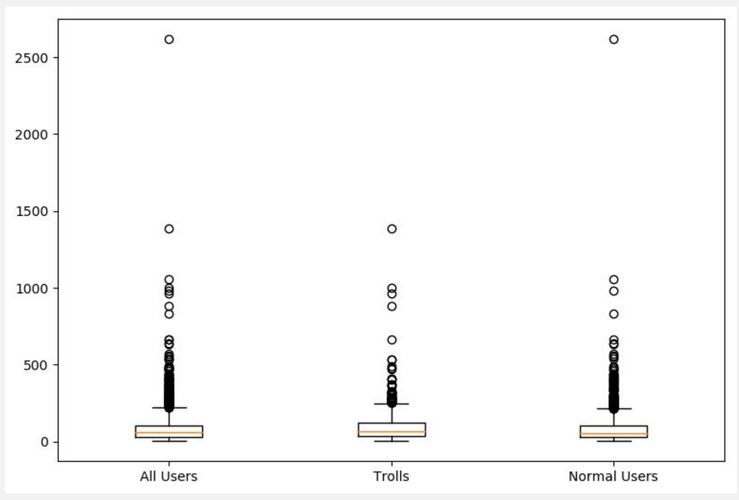


Fig 5.Boxplot of Response Time

Fig 6.Region of user login

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

Conclusion

Building ground truth is the most difficult problem in our work. The user's intention is difficult to define even if we use human eyes to check them 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 model, response time analysis has 0.28 F1-score and 0.17 recall yet 0.97 precision. Combine the three models, 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.

Future work

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 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 future.