A Comprehensive Analysis of Detection of Online Paid Posters

Cheng Chen, Kui Wu, Venkatesh Srinivasan and Xudong Zhang

Abstract We initiate a systematic study to help distinguish a special group of online users, called hidden paid posters, or termed "Internet water army" in China, from the legitimate ones. On the Internet, the paid posters represent a new type of online job opportunities. They get paid for posting comments or articles on different online communities and web sites for hidden purposes, e.g., to influence the opinion of other people toward certain social events or business markets. While being an interesting strategy in business marketing, paid posters may create a significant negative effect on the online communities, since the information from paid posters is usually not trustworthy. When two competitive companies hire paid posters to post fake news or negative comments about each other, normal netizens may feel overwhelmed and find it difficult to put any trust in the information they acquire from the Internet. In this paper, we thoroughly investigate the behavioral pattern of online paid posters based on real-world trace data. We design and validate a new detection mechanism, using both nonsemantic analysis and semantic analysis, to identify potential online paid posters. Our test results with real-world datasets show a very promising performance.

Keywords Online paid posters · Behavioral patterns · Spam detection · Machine learning · Semantic analysis

C. Chen (✉) · K. Wu · V. Srinivasan
University of Victoria, 3800 Finnerty Road, Victoria, Canada
e-mail: cchenv@uvic.ca

K. Wu
e-mail: wkui@cs.uvic.ca

V. Srinivasan
e-mail: venkat@cs.uvic.ca

X. Zhang
Peking University, No. 5 Yiheyuan Road, Haidian District, Beijing, China
e-mail: zhang.xd927@gmail.com

Ö. Ulusoy et al. (eds.), *Recommendation and Search in Social Networks*,
Lecture Notes in Social Networks, DOI 10.1007/978-3-319-14379-8_6

1 Introduction

Working as an online paid poster is a rapidly growing job opportunity for many online users, mainly college students and the unemployed people. These paid posters are referred to as the "Internet water army" in China because of the large number of people who are well organized to "flood" the Internet with purposeful comments and articles. This new type of occupation originates from Internet marketing, and it has become popular with the fast expansion of the Internet. Often hired by public relationship (PR) companies, online paid posters earn money by posting comments and articles on different online communities and web sites. Companies are always interested in effective strategies to attract public attention toward their products. The idea of online paid posters is similar to word-of-mouth advertisement. If a company hires enough online users, it would be able to create hot and trending topics designed to gain popularity. Furthermore, the articles or comments from a group of paid posters are also likely to capture the attention of common users and influence their decision. In this way, online paid posters present a powerful and efficient strategy for companies. To give one example, before a new TV show is broadcasted, the host company might hire paid posters to initiate many discussions on the actors or actresses of the show. The content could be either positive or negative, since the main goal is to attract attention and trigger curiosity.

However, the consequences of using online paid posters are yet to be seriously investigated. While online paid posters can be used as an efficient business strategy in marketing, they can also act in some malicious ways. Since the laws and supervision mechanisms for Internet marketing are still not mature in many countries, it is possible to spread wrong, negative information about competitors without any penalties. For example, two competitive companies or campaigning parties might hire paid posters to post fake, negative news or information about each other. Obviously, ordinary online users may be misled, and it is painful for the web site administrators to differentiate paid posters from the legitimate ones. Hence, it is necessary to design schemes to help normal users, administrators, or even law enforcers quickly identify potential paid posters.

Despite the broad use of paid posters and the damage they have already caused, it is unfortunate that there is currently no systematic study to solve the problem. This is largely because online paid posters mostly work "underground" and no public data is available to study their behavior. We make the following contributions:

1. We collect real-world data from popular web sites regarding a famous social event, in which we believe there are potentially many hidden online paid posters.
2. We statistically analyze the behavioral patterns of potential online paid posters and identify several key features that are useful in their detection.
3. We integrate semantic analysis with the behavioral patterns of potential online paid posters to further improve the accuracy of our detection.

2 Related Work

Previous work focused on forum and blog spammers who posted advertisements or malicious URLs on the web sites. The spammers in those scenarios used software to post malicious comments on their forums and blogs to change the results of search engine or to make their sites popular. However, the definition of spam has been extended to a much wider scope. Basically, any user whose behavior interferes with normal communication or aids the spread of misleading information is specified as a spammer. Examples include comment spammers and review spammers in social media and online shopping stores.

Yin et al. [1] studied so-called online harassment, in which a user intentionally annoyed other users in a web community. They investigated the characteristics of harassment using local features, sentimental features, and contextual features. Gao et al. [2] conducted a broad analysis on spam campaigns that occurred in Facebook network. From the dataset, they noticed that the majority of malicious accounts were compromised accounts, instead of "fake" ones created for spamming. Such compromised accounts can be obtained through trading over a hidden online platform, according to [3].

Ott et al. [4] detected fictitious opinions that are deliberately and *intelligently crafted* to be authentic. To emphasize this point, the authors set strict quality control on the fictitious posts, that is, any submission found to be of insufficient quality, e.g., written for the wrong hotel, unintelligible, unreasonably short, plagiarized, etc., will be rejected. This problem is different from ours, since we do not focus on the *deceptive* opinions, but instead we aim at detecting *disruptive* comments, which are not hard to determine if a person has enough resource and time, i.e., she/he has collected a large pool of comments from different sites, a large pool of user IDs, and she/he has enough patience to read all comments and compare the comments from a same user.

The work by Jindal and Liu [5] is close to our research. They studied a dataset crawled from Amazon.com and tried to detect "opinion spam" or "review spam." In [5], the authors assumed the review spammer acts individually. In recent work [6, 7], the authors focused on detecting groups of spammers. They found that labeling groups of spammers were easier because the behavior of a group of spammers could be detected if the spammers had similar behavior when they wrote reviews for products. Our case study, however, *largely differs* from those in [5–7]. As demonstrated in this paper, paid posters involved in business conflicts have different posting patterns and do not exhibit the features presented in [5–7]. In our work, the data is not reviews for products, but any social comments, which are *shorter than the reviews* in general, regarding various aspects of the two companies, including for example the chairman, the products, and the marketing activities. As a result, the features used in our work are different from those in [5–7]. Furthermore, our semantic analysis method to improve detection performance is based on the identification of common content words and is different from those in [5–7].

3 Data Collection and Manual Labeling

3.1 Data Collection

In this paper, we analyze a business dispute between 360 and Tencent, two IT companies. [1] We collected news reports and relevant comments regarding this special social event from two famous Chinese news web sites: Sina.com [8] and Sohu.com [9]. *Sina dataset* and *Sohu dataset* will be used as the training data and test data for our detection model, respectively. We searched all the news reports and comments from Sina.com and Sohu.com over the time period from September 10, 2010 to November 21, 2010. As a result, we found 22 news reports in Sina.com and 24 news reports in Sohu.com. For each comment of each news report, we recorded the following relevant information: *Report ID*, *Sequence No.*, *Post Time*, *Post Location*, *User ID*, *Content*, and *Response Indicator* (i.e., whether the comment is a new comment or a reply to another comment).

3.2 Manual Identification

In order to analyze the behavioral pattern and classify potential paid posters and normal users, we need to find out the "ground truth" in the two datasets and we use the following guidance:

1. Users who post meaningless or contradicting comments. For example, the comments are not even slightly related to the topic in discussion. Also, a user may post multiple comments showing completely different opinions.
2. Users who post many short comments without any supporting evidence. For example, short comments like "I like 360" and "360 is good" are less likely from reasonable users involved in serious discussion.
3. Users who post negative and irrational comments to attack other persons.
4. Users who post multiple duplicate or near duplicate comments. Unlike the above three behaviors, we do not consider it as a critical criterion in labeling the datasets because both potential paid posters and normal users can have this behavior. Before making final decision, users with this behavior are carefully considered together with other criteria.

We are confident about our labels, as we believe any reasonable person would agree that a user who posts seven "I hate 360" within 2 minutes should be a potential paid poster; and any reasonable person would also agree that a user who posts both "I really like 360 because it protects my computer so well" and "It is really bad that

[1] For a full description of this dispute, please refer to http://en.wikipedia.org/wiki/360_v._Tencent.

360 steals my private information. I hate 360" should be a potential paid poster. As a result, 70 and 82 potential paid posters were identified from the Sina dataset and the Sohu dataset, respectively.

Remark 1 Finding the "gold standard" ground truth is still an open problem and no research has been able to solve this problem. Existing efforts use cross-checking among multiple annotators, as what we have done in this work. One extreme way is to hire paid posters to post fake comments and collect the corresponding texts. This method was used by Ott et al. [4], who worked on a related (but different) problem and obtained "gold standard" labels by using Amazon Mechanical Turk (AMT) to hire turkers to post fictitious hotel reviews. Nevertheless, even with such a costly method, it is difficult to obtain "gold standard" labels, because they have no guarantee that posts not from their hired tuckers are truthful. Due to the above reason, we use the word *potential* to avoid the nontechnical argument about whether a manually selected paid poster is really a paid poster. Any absolute claim is not possible unless a paid poster admits to it or his employer discloses it, both of which are unlikely to happen. The lack of "gold standard" is common in social studies, although it has been criticized and not understood by many engineers.

4 Nonsemantic Analysis

In this section, we perform statistical analysis to investigate objective features that are useful in capturing the potential paid posters' special behavior. We use Sina dataset as our training data and thus we only perform statistical analysis on this dataset. We mainly test the following four features: percentage of replies, average interval time of posts, the number of days the user remains active and the number of news reports that the user comments on. In the following figures, we use "pp" and "nu" to denote potential paid posters and normal users, respectively.

1. *Percentage of Replies*. In this feature, we calculate the probability whether a user tends to post new comments or reply to others' comments. We conjecture that potential paid posters may not have enough patience to read others' comments and reply. Therefore, they may create more new comments. Figure 1 shows the statistical results, with respect to the density and cumulative density function of reply ratio.
2. *Average Interval Time of Posts*. We calculate the average interval time between two consecutive comments from the same user. Note that it is possible for a user to take a long break (e.g., several days) before posting messages again. To alleviate the impact of long break times, for each user, we divide his/her active online time into epochs. Within each epoch, the interval time between any two consecutive comments cannot be larger than 24 h. We calculate the average interval time of posts within each epoch, and then take the average again over all the epochs. Figure 2 shows the statistical results for the probability distribution of interval posting time.

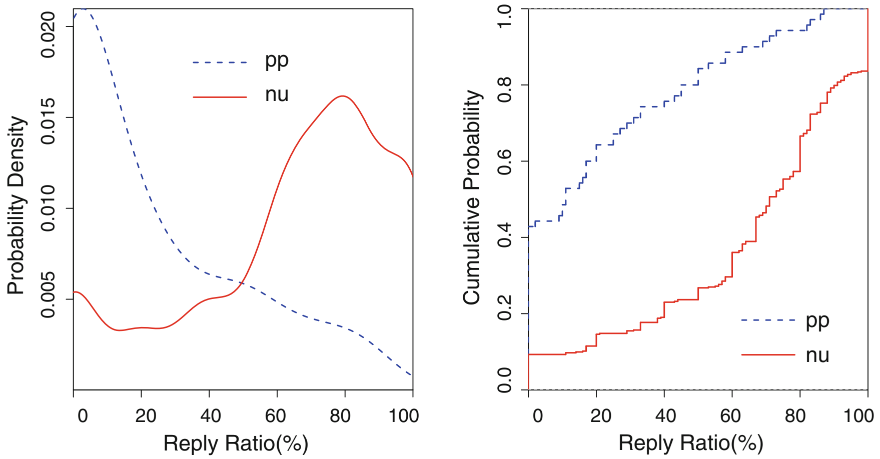


Fig. 1 The PDF and CDF of reply ratio

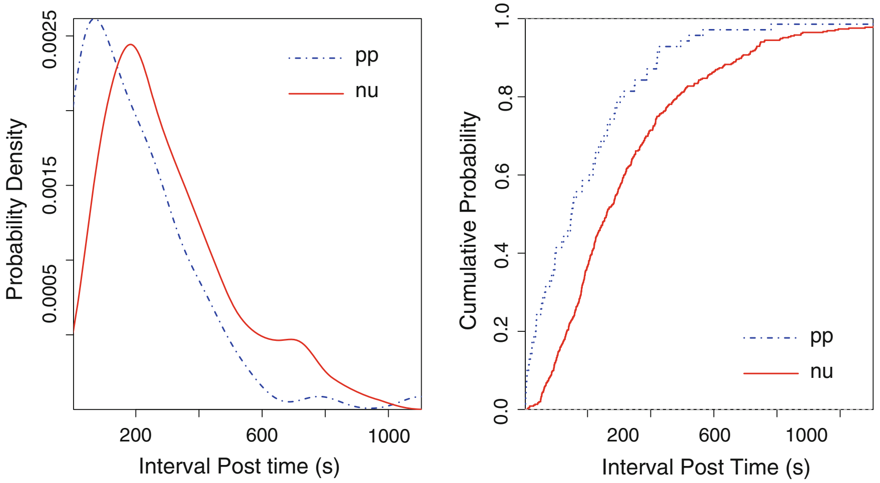


Fig. 2 The PDF and CDF of average interval time

3. *Active Days*. We analyze the number of days that a user remains active online. This information can be extracted from the time stamp of their comments. We divide the users into 7 groups based on whether they stay online for 1, 2, 3, 4, 5, 6 days and more than 6 days, respectively. Potential paid posters usually do not stay online using the same user ID for a long time. Once a mission is finished, a paid poster normally discards the user ID and never uses it again. When a new mission starts, a paid poster usually uses a different user ID, which may be newly created or assigned by the *resource team*. Figure 3 shows the statistical result. In the figures, “7” at the x-axis is the number of active days for 7 days or more.

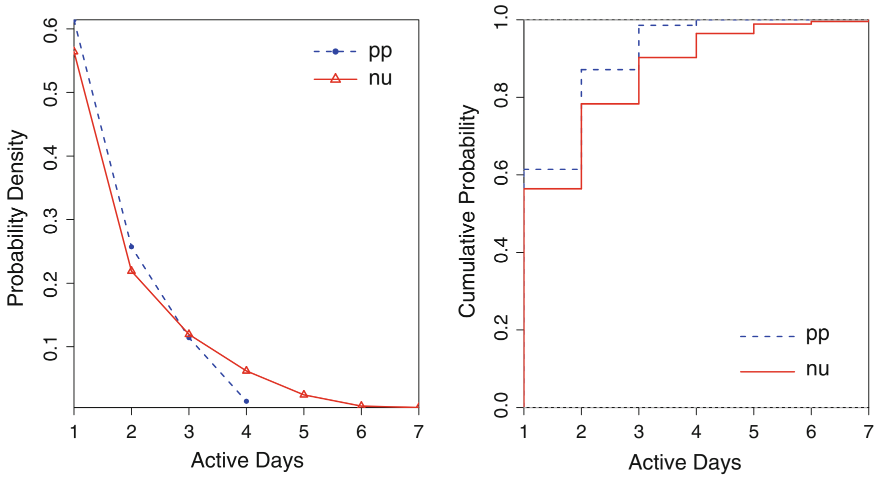


Fig. 3 The PMF and CDF of number of active days

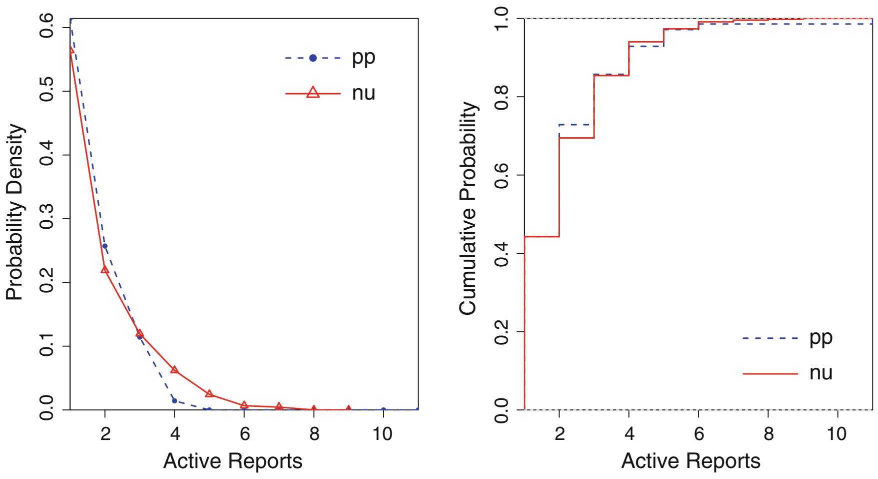


Fig. 4 The PMF and CDF of number of active news reports

4. *The Number of News Reports*. We study the number of news reports for which a user has posted comments. Both Sina and Sohu have nearly 20 news reports. Figure 4 shows the corresponding graphs.

We can derive the following conclusions from Figs. 1, 2, 3 and 4:

1. Potential paid posters tend to have smaller reply ratio.
2. Potential paid posters only care about finishing their jobs as soon as possible and do not have enough interest to get involved in the online discussion. 60 % potential paid posters post comments within interval time of 200 seconds.

3. Potential paid posters are not willing to stay for a long time. They instead tend to accomplish their assignments quickly and once it is done, they would not visit the same web site again.
4. Potential paid posters and normal users have similar distribution with respect to the number of commented news reports. This indicates that the number of commented news reports alone may not be a good feature for the detection of potential paid posters.

5 Semantic Analysis

An important criterion in our manual identification of a potential paid poster is to read his/her comments and make a choice based on common sense and online experience. While it is hard to design a detection system that understands the meaning of a comment, we observed that potential paid posters tend to post similar comments on the web. In many cases, a potential paid poster may copy and paste existing comments with slight changes. This provides the intuition for our semantic analysis technique.

Our basic idea is to search for similarity between comments. To do this, we first need to overcome the special difficulty in splitting a Chinese sentence into words and phrases. We used a famous Chinese splitting software, called ICTCLAS2011 [10], to cut a sentence into words. It translates a sentence into a list of content words. For a given pair of comments, we compare the two lists of content words. As mentioned before, a paid poster may make slight changes before posting two similar comments. Therefore, we may not be able to find an exact match between the two lists. We first find their common content words, and if the ratio of the number of common content words over the length of the shorter content word list is above a threshold value (e.g., 80 % in our later test), we conclude that the two comments are similar. If a user has multiple pairs of similar comments, the user is considered a potential paid poster. Note that similarity of comments is not transitive in our method.

We found that a normal user might occasionally have two *identical* comments. This may be caused by the slow Internet access, due to which the user presses the *submit* button twice before his/her post is displayed. Our manual check of these users confirmed that they are normal users, based on the content they posted. To reduce the impact of the "unusual behavior of normal users," we set the threshold of similar pairs of comments to 3. This threshold value is demonstrated to be effective in addressing the above problem.

We performed the semantic analysis over the Sina dataset. The result is shown in Fig. 5.

In the figure, "6" on the x-axis means the number of similar pairs is larger than or equal to 6. The two groups of users obviously show different patterns. Normal users have much higher probability to post different comments. In the opposite, the potential paid posters have many similar pairs of comments in their profiles. Therefore, it is important to monitor the number of similar pairs of comments in a user's profile as it is a significant indication of malicious behavior.

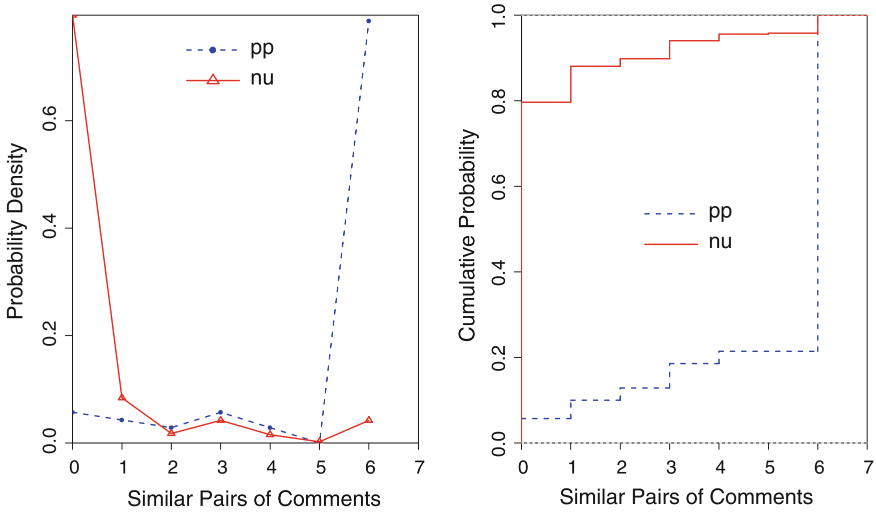


Fig. 5 The PMF and CDF of the number of similar pairs of comments

6 Classification

The objective of our classification system is to classify each user as a potential paid poster or a normal user using the features investigated in Sects. 4 and 5. According to the statistical and semantic analysis results, we found that any single feature is not sufficient to locate potential paid posters. Therefore, we compare the performance of different combinations of the five features discussed in the previous two sections in our classification system. We model the detection of potential paid posters as a binary classification problem and solve the problem using a support vector machine (SVM) [11].

We used LIBSVM [12] as the tool for training and testing. By default, LIBSVM adopts a radial basis function [11] and a 10-fold cross-validation method to train the data and obtain a classifier. We did not tune any model parameter of libsvm and liblinear. All results came from the default settings, so that we could compare the results in a general way. The Sina dataset is divided into 10 subsets of equal size. Then the model is trained on the 9 subsets and tested on the remaining subset. The process returns a model with the highest cross-validation accuracy. After training the classifier with the Sina dataset, we used the classifier to test the Sohu dataset.

We evaluate the performance of the classifier using the four metrics: precision, recall, f measure and accuracy, defined as follows:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

$$\text{F measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{True Negative} + \text{True Positive}}{\text{Total Number of Users}} \tag{4}$$

6.1 Classification Without Semantic Analysis

To simplify the notation, the five features, *reply ratio*, *average interval posing time*, *active days*, *active reports*, and *degree of similarity* are labeled as features “1”, “2”, “3”, “4” and “5,” respectively. The first four features are statistical ones while the last is a semantic feature.

We firstly focus on the classification only using statistical analysis results based on the four statistical features. Different combinations are applied to test their performance for identification. We train the SVM model using the Sina dataset with different combinations of the features. Then we test the model with the Sohu dataset to see the performance. Note that combinations that result in 0 true positive and 0 false positive are not considered. The results are shown in Fig. 6.

Although the (1-2)-feature test has the highest precision, its recall and f measure are very low, showing that the (1-2)-feature can hardly separate different classes of users. This result suggests that the first two features lead to significant bias and we need to add more features to our classifier. With features 3 and 4 considered, we observe better performance. For example, the (1-2-3)-feature test has better performance over all the metrics, except precision.

Nevertheless, we notice when we use only nonsemantic features to train the SVM model, the overall performance on the four metrics is not good enough to claim acceptable performance. Particularly, the low precision and accuracy results indicate that the SVM classifier using the four nonsemantic features as its vector set is unreliable and needs to be improved further. We achieve this by adding the semantic analysis to our classifier.

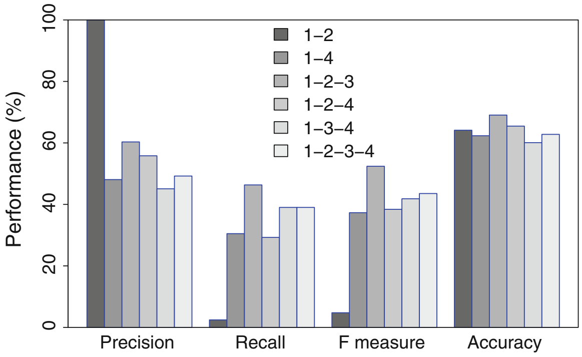
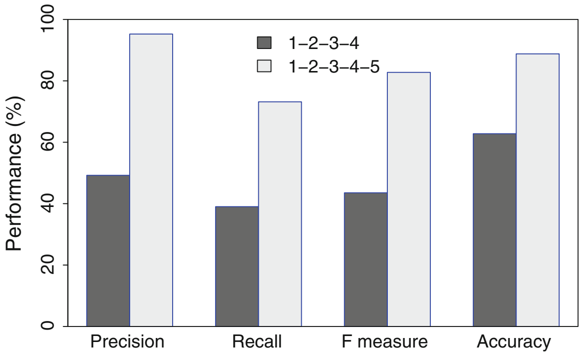


Fig. 6 The performance of different combinations of statistical features

Fig. 7 The performance of statistical and semantic features



6.2 *Classification with Semantic Analysis*

As described in Sect. 5, we have observed that online paid posters tend to post a larger number of similar comments on the web. Based on this observation we have designed a simple method for semantic analysis. We test the performance of all the five features. After integrating this semantic analysis method into our SVM model, we observed the much improved performance results as shown in Fig. 7.

The results clearly demonstrate the benefit of using semantic analysis in the detection of online paid posters. The precision, recall, f measure, and accuracy have been improved to 95.24, 73.17, 82.76 and 88.79 %, respectively. Based on these results, the semantic feature can be considered as a useful and important supplement to other features. The reason why the semantic analysis improves performance is that online paid posters often try to post many comments with some minor changes on each post, leading to similar sentences. This helps the paid posters post many comments and complete their assignments quickly, but also renders it easy to detect them.

Having shown that the five proposed features together lead to higher performance, we add additional tests with random forest [13]. Random forest consists of multiple decision trees whose prediction is easier to explain (i.e., the criteria for making predictions) and no parameter tuning is required. In the learned tree structures, branches are conjunctions of features that lead to class labels which are represented by leaves. In practice, it will aid better interpretation of decision-making. The performance of SVM and random forest is shown in Fig. 8. The result shows that random forest does not perform as well as SVM.

6.3 *Classification Using only Text Information*

As a comparison to the previous method, we use a typical information retrieval approach to identify potential paid posters in this subsection. We now use only text information (individual words in comments) for training the classifier. Specifically, we treat each user's comments as an individual document and it becomes a binary

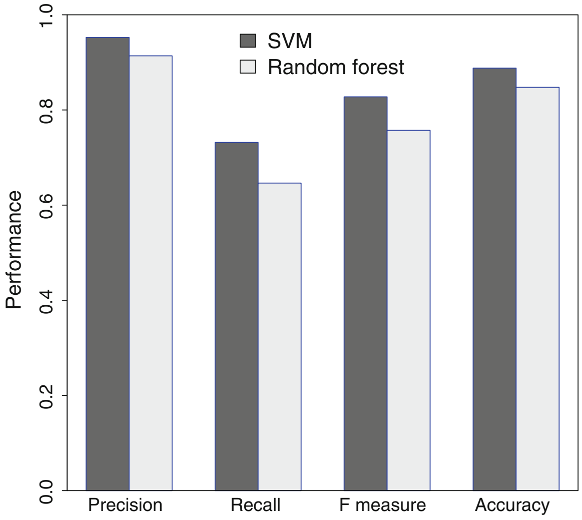


Fig. 8 Performance comparison of SVM and random forest

document classification problem; to detect potential paid posters is to classify each document into two distinct groups (i.e., malicious and normal).

6.3.1 Feature Selection

We use the Chi-square method [14, 15] to retrieve a bag of feature words, a standard methodology of extracting features in documentation classification.

We define variables A, B, C, and D in Table 1. For example, A is the number of paid posters who have a specific word in the comments. D is the number of normal users who do not have the specific word.

After we collect the statistic information for every individual word, we can then compute Chi-square values. The Chi-square value of a word in the document collection is defined as

$$\text{chisquare(word, classification)} = \frac{(AD - BC)^2}{(A + B)(C + D)} \tag{5}$$

We compute the Chi-square value for each word in the training document collection, sort them in descending order and retrieve the first d words as the bag of the most predictive features.

Table 1 Chi-square feature selection

Feature selection	Paid posters	Normal users	Total
Has word	A	B	$A + B$
Not word	C	D	$C + D$
Total	$A + C$	$B + D$	N

6.3.2 Vectorization

After selecting feature words from the document collection, we can then vectorize each document by associating it with a vector of dimension d. We compute the weight for each dimension using the TF/IDF approach [16].

6.3.3 Classifier

To study the performance of text information for this document classification problem, we explore different nonlinear classifiers as well as linear ones on our dataset and compare their prediction results. In the following, we use Liblinear [17] for the linear classifier.

Compared to the general-purpose SVM solver Libsvm, Liblinear is exclusively used for linear classification, i.e., it supports logistic regression and linear support vector machines. Without using kernels, Liblinear can train a much larger set via a linear classifier. Consequently, Liblinear is considered a better choice over Libsvm when handling large-scale datasets (e.g., document classification) for which using nonlinear mappings does not provide additional benefit.

Tables 2 and 3 list candidate models to be tested in the experiment.

6.3.4 Performance Evaluation

In order to show the impact of different dimensions, we use four settings for the following tests, i.e., $d = 100$, $d = 200$, $d = 300$, and $d = 400$. The results are shown in Fig. 9.

Table 2 Libsvm kernel types

Kernel type	Description
t0	Linear: $u' * v$
t1	Polynomial: $(\gamma * u' * v + \mathrm{coef0})^{\circ}$
t2	Radial basis function (RBF): $\exp(-\gamma * \lvert u - v \rvert^2)$
t3	Sigmoid: $\tanh(\gamma * u' * v + \mathrm{coef0})$

Table 3 Liblinear solver types

Solver type	Description
s0	L2-regularized logistic regression (primal)
s1	L2-regularized L2-loss support vector classification (dual)
s2	L2-regularized L2-loss support vector classification (primal)
s6	L1-regularized logistic regression

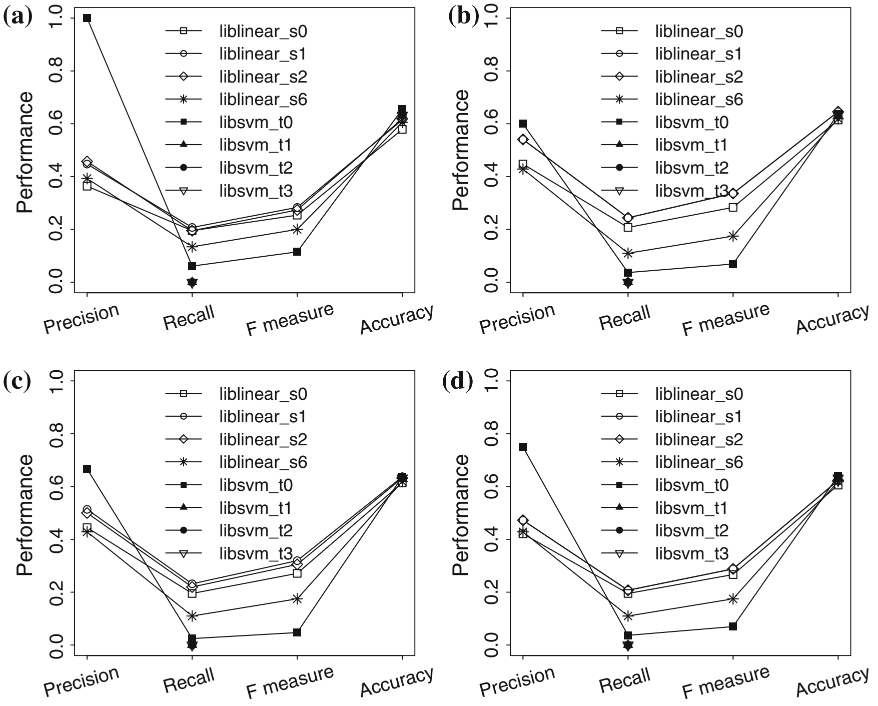


Fig. 9 Performance curves for different dimensions for Libsvm and Liblinear **a** $d = 100$ **b** $d = 200$ **c** $d = 300$ **d** $d = 400$

From Fig. 9 we observe that models trained by Liblinear have better performance over the ones trained by Libsvm. Specifically, Liblinear models of $d = 200$ have the overall best performance. Recall and f measure of Liblinear are significantly higher than Libsvm, even if precision of Libsvm with t0 is the highest. Note that metrics of precision and f measure for Libsvm models with nonlinear kernels (t1, t2, and t3) in the figures are not available (corresponding to the missing points in Fig. 9), because those models only return negative predictions. It indicates that nonlinear SVM classifiers are not valid in this high-dimensional classification problem. All valid models have similar accuracy measurement.

In addition, an interesting observation is that the best Liblinear model does not exceed the performance of Libsvm model trained by (1-2-3-4)-feature, which is described in previous subsections. The reason is that the high-dimensional feature space is too sparse to facilitate the learning algorithm. The sparsity is due to the fact that a user's comments tend to be short and the selected feature words cannot provide enough coverage even if we group each user's comments.

In order to evaluate the performance over all features mentioned in this paper, we add 200-dimension text information into the feature space, labeled by (1-2-3-4-5). We then use Liblinear to train a linear classifier and evaluate it over the Sohu test set. The results are shown in Figs. 10 and 11.

Fig. 10 Liblinear combination of features 1-2-3-4-5 and 200-dimension text feature

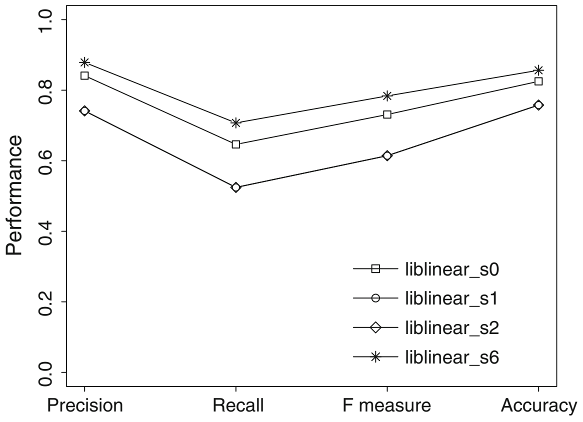


Fig. 11 Performance comparison with/without 200-dimension text feature

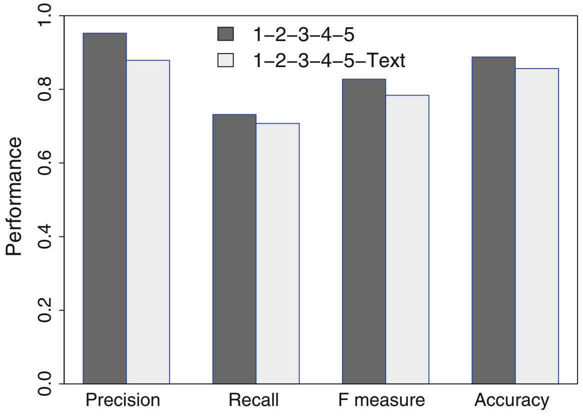


Figure 10 shows that Liblinear with s6 outperforms other Liblinear models. In Fig. 11, we compare the best results of Liblinear with s6 (including 200-dimension text feature) to the previous one (Libsvm performance excluding text feature). It shows that adding 200-dimension text feature would unfortunately harm the overall performance.

6.4 Classification Using Unsupervised Learning

For unsupervised learning, we firstly merged Sina dataset and Sohu dataset and applied K-means clustering algorithm to obtain K clusters. If the five features have the ability to distinguish paid posters from normal users, we expect that paid posters should be grouped into a cluster. In our work, we only need two clusters, one for paid posters and one for normal users. Furthermore, to check the reliability of our features, we studied two more cases, corresponding to $K = 3$ and $K = 4$.

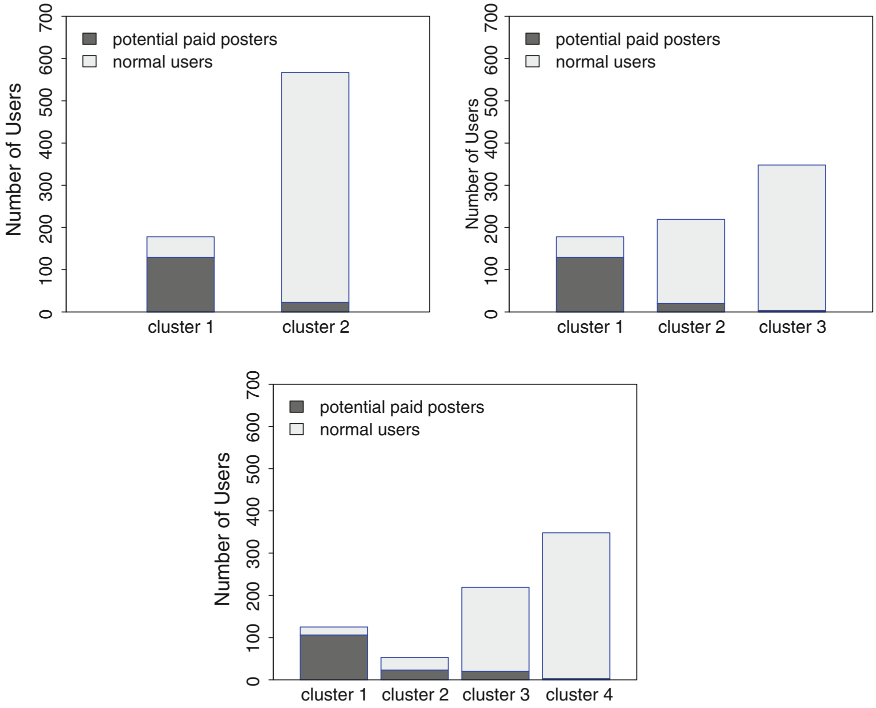


Fig. 12 Clustering analysis $K = 2, 3, 4$

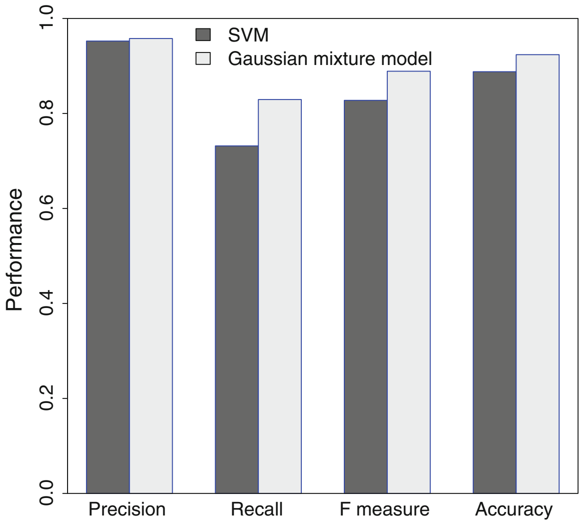
Figure 12 shows the size of each cluster as well as the number of potential paid posters and normal users in each cluster.

From the figures, we notice that when $K = 2$, a large proportion (approximately 85 %) of potential paid posters is assigned to a particular cluster (cluster 1). When $K = 3$ and $K = 4$, cluster 1 (the group of paid posters) remains stable. Nevertheless, the other cluster (the group of normal users) is further divided into smaller clusters. This phenomenon suggests that although normal users might have different behavioral patterns, they in general behave much different from potential paid posters.

We also notice that a small number of normal users are assigned to cluster 1. This is because our manual labeling uses human intelligence (refer to Sect. 3), which cannot be completely captured by the five features. This poses the challenge of developing more intelligent detection mechanism for our future work.

We now compare the clustering model with the supervised model. We train a Gaussian mixture model [18], a generalization of K-means, on Sina dataset and test it on Sohu dataset. It incorporates information about the means ($\boldsymbol{\mu}$) and covariance matrix ($\boldsymbol{\Sigma}$) of features of the data. As an unsupervised approach, it applies expectation–maximization algorithm [19] to estimate the model parameters. Its prediction is based on the sample's probability of being assigned to each cluster. We compare its performance with that of SVM in Fig. 13.

Fig. 13 Performance comparison of SVM and Gaussian mixture model



In Fig. 13, we can see that Gaussian mixture model outperforms SVM in all metrics, even the lowest recall measure exceeds 80 %. The result implies that not all features of the data satisfy the assumption of independent and identical distribution. In addition, the test using Gaussian mixture model also demonstrates the effectiveness of the five proposed features to differentiate the malicious from the normal.

7 Conclusions and Future Work

Detection of paid posters behind social events is an interesting research topic and deserves further investigation. In this paper, we disclose the organizational structure of paid posters. We also collect real-world datasets that include abundant information about paid posters. We identify their special features and develop effective techniques to detect them. The performance of our classifier, with integrated semantic analysis, is quite promising on the real-world case study, as confirmed in both supervised learning and unsupervised learning techniques.

This work is our preliminary effort to battle online paid posters. It requires a prolonged and systematic effort to reach a complete solution, as the online paid posters evolve continuously and present new challenges to the detection mechanism. We will further improve our detection system and evaluate the system in a broader and larger dataset. We wish this work would attract further research activities.

Acknowledgments We thank Natural Sciences and Engineering Research Council of Canada (NSERC) and Mathematics of Information Technology And Complex Systems (MITACS) for the funding support. We thank MIT Tech. Review [20] for announcing our work to the public.

References

1. Yin D, Xue Z, Hong L, Davison B, Kontostathis A, Edwards L (2009) Detection of harassment on web 2.0. In: Proceedings of the content analysis in the web 2
2. Gao H, Hu J, Wilson C, Li Z, Chen Y, Zhao BY (2010) Detecting and characterizing social spam campaigns. In: ACM conference on computer and communications security, pp 681–683
3. Staff E (2010) Verisign: 1.5 m facebook accounts for sale in web forum. http://www.pcmag.com/article2/0,2817,2363004,00.asp
4. Ott M, Choi Y, Cardie C, Hancock JT (2011) Finding deceptive opinion spam by any stretch of the imagination. In: ACL, pp 309–319
5. Jindal N, Liu B (2008) Opinion spam and analysis. In: Proceedings of the international conference on web search and web data mining, WSDM'08, ACM, New York, pp 219–230
6. Mukherjee A, Liu B, Glance NS (2012) Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st international conference on World Wide Web: pp 191–200
7. Mukherjee A, Liu B, Wang J, Glance NS, Jindal N (2011) Spotting fake reviewer groups in consumer reviews. In: WWW: pp 191–200
8. Sina.com. www.sina.com.cn Accessed Jan 2011
9. Sohu.com. www.sohu.com Accessed Jan 2011
10. ICTCLAS2011. http://hi.baidu.com/drkevinzhang/home Accessed Mar 2011
11. Cristianini N, Shawe-Taylor J (2006) An introduction to support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge
12. Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology, vol **2**(27), pp 27:1–27:27 Software available at http://www.csie.ntu.edu.tw/.
13. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
14. Zheng Z, Wu X, Srihari R (2004) Feature selection for text categorization on imbalanced data. SIGKDD Explor Newsl 6(1):80–89
15. Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305
16. Joachims T. (1997) A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In: Proceedings of the fourteenth international conference on machine learning ICML'97, Morgan Kaufmann, San Francisco, pp 143–151
17. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9:1871–1874
18. Marin JM, Mengersen K, Robert CP (2005) Bayesian modelling and inference on mixtures of distributions. Handb Stat 25:459–507
19. Dempster AP, Laird NM, Rubin DB et al (1977) Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc 39(1):1–38
20. MIT-Tech-Review. www.technologyreview.com/blog/arxiv/27357/ Accessed Nov 2011