Paid Review and Paid Writer Detection

Man-Chun Ko
National Taiwan University
Department of Computer Science and Information Engineering
Taipei, Taiwan
mcko@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
National Taiwan University
Department of Computer Science and Information Engineering
Taipei, Taiwan
hhhuang@nlg.csie.ntu.edu.tw

Hsin-Hsi Chen
National Taiwan University
Department of Computer Science and Information Engineering
Taipei, Taiwan
hhchen@ntu.edu.tw

ABSTRACT

There has been a surge in opinion-sharing in the public domain. Some opinions greatly influence our decisions, e.g., the choice of purchase. Malicious parties or individuals exploit social media by generating fake reviews for opinion manipulation. This paper aims to investigate the phenomenon of online paid restaurant reviews by bloggers. Our research provides an insight into some characteristics of paid reviews and their authors. We then explore a set of features based on our observations and detect paid reviews and paid bloggers using supervised machine learning techniques. Experimental results show the effectiveness of our approach.

CCS CONCEPTS

• Computing methodologies→Artificial intelligence→ Natural language processing •Applied computing → Law, social and behavioral sciences

KEYWORDS

Opinion Spam; Fake Review; Paid Review; Blogging

ACM Reference format:

M.-Ch. Ko, H.-H. Huang, and H.-H. Chen. 2017. Paid Review and Paid Writer Detection. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017*, 9 pages.
http://dx.doi.org/10.1145/3106426.3106433

1 INTRODUCTION

With the advent of the Internet, the general population became adapted at sharing and exchanging their opinions on the web, and these opinions can subsequently affect people's thoughts and decisions. Since online opinions play a major role in consumers' decisions, they give merchants strong motivations to manipulate their reputations on the Internet. Malicious individuals or parties, namely *opinion spammers,* are involved to promote products or political candidates by publishing dishonest reviews, and such deceptive information might mislead potential customers or voters. Opinion spam has attracted significant attention from both business and research communities.

This paper addresses the issue of paid review detection in restaurant reviews. Customers usually complain about the gap between reviews and the real dining experience from time to time. One of the reasons is that those reviews are written by the bloggers receiving payment or free food/service from restaurants. Deceitful restaurants may provide better food or service to paid writers, so the dining experience of writers is better than that of customers. Besides, restaurants are likely to ask writers praise their meals using the overstated or unfair content.

Most previous works investigated fake review detection in online restaurant rating websites such as yelp.com and dianping.com. In this work, our material is the reviews on blogs. Blogging is popular for people to share their daily experiences or publish their opinions and emotions. Compared to restaurant rating websites, there are almost no restrictions on publishing reviews on blogs. Writers (bloggers) manipulate the contents and the advertising on their blogs by their own. Restaurant reviews on blogs are more writer-centric, and bloggers, especially the paid ones, care about their reputation. Thus, some bloggers attempt to hide the fact that they are paid. In this paper, the special aspects of blogging are analyzed in detail.

We collect the restaurant reviews from Pixnet (www.pixnet.net), the largest blog platform in Taiwan, which has over 4.5 million users. Naturally, Pixnet's popularity intrigues opinion spammers to promote their targets. These testimonials or reviews on Pixnet cover a variety of targets such as consumer electronics, cosmetics, and restaurants. Some honest paid writers on Pixnet claim their reviews being sponsored, but many spammers carefully hide such information. The lack of ground-truth is a crucial problem in opinion spam researches. Fortunately, Professional Technology Temple (PTT), the largest online bulletin board in Taiwan, can provide reliable ground truth. Some sub-forums like the "Makeup" board have completely forbidden the commercial posts, and some sub-forums force authors to specify their business relationship if they cooperate with others. In the "Food" board, which has over 100,000 restaurant reviews, authors have obligated to label their

WI '17, August 23-26, 2017, Leipzig, Germany

ACM ISBN 978-1-4503-4951-2/17/08…$15.00
http://dx.doi.org/10.1145/3106426.3106433

commercial posts since 2012. It is difficult to identify deceitfulness for readers, but with the labels from PTT, we can study such covert marketing campaigns.

In this study, we call our detection targets *paid review* and *paid writer* rather than opinion spam and opinion spammer. Compared to the traditional opinion spammers, who make up the fake reviews, most of the paid writers in our dataset have practical dining experiences. Contents of paid reviews include not merely personal experiences, but more words of flattery to influence customers' decisions. We aim to find out characteristics of these paid reviews and paid writers, and propose models to spot the paid reviews and paid writers automatically.

The contributions made in this study are three folds:

1. We collect the restaurant reviews from Pixnet and ground truth from PTT to build a novel dataset. To the best of our knowledge, this is the first attempt to study the opinion spam activity on blogs.
2. The behavior of paid writers and the contents of paid reviews are analyzed to identify their characteristics. A set of features extracted from contents and writer behavior is proposed to detect paid reviews and paid writers.
3. We conduct a set of experiments using extracted features and learning-based techniques to detect paid reviews and paid writers. Experimental results show that both detection models outperform the baselines significantly even in extremely imbalanced cases.

1 RELATED WORK

Jindal and Liu [6] are the first researchers to study opinion spam. They crawled product reviews from Amazon, analyzed spam activities, and detected fake reviews. In recent years, the research community has put significant effort in opinion spams and explored different dimensions of opinion spams.

Due to the subtlety nature, ground truth acquisition is difficult. Researchers collected the ground truth in different ways and built several datasets to address the problem of opinion spams. Amazon has attracted many researchers to study opinion spam detection [3] [6] [8] [13]. To date, people are increasingly using opinion-sharing websites. Mukherjee et al. [14] studied the opinion spams in Yelp, a popular website providing local search, ratings, and reviews services. They also obtained the ground truth from Yelp filter, and claimed that the ground truth is sufficiently reliable since Yelp filter probably uses some internal metrics (e.g., IP addresses and user logs).

People also like to share their experiences and opinions on web forums. Some malicious companies launched shady campaigns to affect users' decisions. Chen and Chen [1] [2] collected a dataset about a covert marketing campaign on Mobile01, a consumer electronics forum in Taiwan. They built a dataset based on a set of internal records disclosed by a hacker. Ko and Chen [7] scraped a dataset from PTT to analyze the behaviors of a group of cyber army whose goal is to influence voters' decisions for an election campaign.

Spam and spammer detection are often considered as a supervised classification problem. Various machine learning techniques and features have been explored. Most of the opinion spam researches take Support Vector Machine (SVM) with n-grams and behavior features as a baseline. Besides bag of n-grams, other content-based features are extensively used in previous researches. Jindal and Liu [6] derived various features such as the length of the review title and body. Ott et al. [15] improved the performance on their AMT fake reviews with Linguistic Inquiry and Word Count [16], a well-known dictionary which contains 80 psycholinguistic meanings to 4,500 keywords. Because creating multiple fake reviews with different content is cost- and time- consuming, spammers tend to copy the text of existing reviews and publish identical or similar content. Thus content similarity is often applied to spammer [11] and spammer group [13] detection.

Aside from features related to contents, researchers also attempted to extract features from user behaviors. Jindal and Liu [6] proposed features such as the number of helpful feedbacks and rating of the reviews. Other information like product and reviewer information was also explored. Some researchers even leveraged spatial [10] and temporal [18] information for detection. Chen and Chen [1] inspected spammers' behavior on the web forum. They discovered that spammers are the more prolific posters, and they tend to post spam messages during work time rather than leisure time.

2 Dataset

This section introduces our dataset. We first describe how to collect ground truth and reviews from different sources. Then we provide the basic information of our dataset. Lastly, a comparison of the previous opinion spam datasets is presented.

2.1 Data Collection

2.1.1 Data Source

Professional Technology Temple (PTT), the largest online bulletin board in Taiwan, contains over 20,000 boards covering a multitude of topics. The Food board is one of the most popular boards. It has over 100,000 articles, and more than 2,000 articles are submitted monthly. Many people write restaurant reviews on their blogs and post to the Food board to attract more readers. A review on PTT shares similar textual content with its counterpart on the blog, while the blog version usually offers enriched text with photos.

Users on the Food board have noticed that there are too many untruthful reviews manipulated by professional writers. Thus, the administrators of the Food board established a rule to protect the readers since 2012. The author should label her/his article as "promotional" if s/he has business cooperation. Many paid reviewers follow this rule although they may not reveal the commercial background on their blogs. We consider the "promotional" labels as ground-truth of restaurant reviews.

On our scraped data, 70% of reviews on the Food board are hosted on Pixnet, the largest blog platform in Taiwan. The API[2] provided by Pixnet offers metadata such as the pageviews of a review and the subscriber number of a blog. This work focuses on the reviews on Pixnet with the ground truth from PTT.

2.1.2 Data Acquisition

The following shows the whole process of data collection.

1. Scrap all articles from the Food board, and select the articles posted between January 2013 and October 2015.
2. Annotate promotional articles from restaurants. The articles posted by the restaurant owners or employees are removed from our dataset.
3. For each article on the PTT Food board, parse its HTML content with BeautifulSoup[3] and search for all the possible Pixnet URLs using the pattern (http://{user id}.pixnet.net/blog/post/{article id}).
4. Use Pixnet API to download the articles by the URLs from the previous step.
5. If PTT article has only one Pixnet URL, match the Pixnet data we download in the previous step and save both Pixnet data and PTT label to the final review set. Otherwise, calculate similarities of the content belong to each URL, take the most similar one and save it into the final review set.
6. Retrieve information of all bloggers in the final review set, scrap the blog HTML contents for all bloggers and save them into the final writer set.
7. Perform Chinese segmentation on the introduction of each blogger, and on content and title in each review.

2.2 Basic Attributes and Statistics of Dataset

A large portion of the datasets used in previous studies is collected from e-commerce or rating websites such as Amazon or Yelp, while our dataset is scraped from Pixnet, a blog platform. The contents are mostly written in Traditional Chinese, with some English phrases. Moreover, Pixnet API brings in a variety of statistics and metadata. These data are valuable for analysis and even help gain better detection results. Scraped attributes of review and blogger are shown in Tables 1 and 2.

We define the writers who have posted at least one paid review as the *paid writers*, those who never posted paid reviews are called *benign* writers, and the unpaid reviews are called *genuine* reviews. Table 3 shows the basic counts of reviews and writers. Figure 1 shows the number of reviews in each month from January 2013 to August 2015, and Figure 2 illustrates the number of paid reviews each month. Obviously, the number of reviews was increasing since 2013 but stay at the same level after 2014. By contrast, the number of paid reviews is still increasing. Catering industry understands the importance of the word-of-mouth effect. They start leverage customer reviews to promote their food or services.

Table 1. Attributes of reviews

Attribute	Description
address	address provided by the blogger
body	HTML body of the blog post
category	category of the blog post
comment_count	number of the comments in the blog post
cover	URL of the cover photo
hits_daily	daily number of hits
hits_total	total number of hits
id	blog post id
images	URLs of images in the blog post
is_top	whether this post is pinned on the top of the blog
link	link of the blog post
public_at	post time of the blog post
sns_facebook/ plurk/twitter	whether this post is shared to Facebook, Plurk, and Twitter
tags	hashtag used in the blog post
thumb	thumbnail picture URL
title	title of the blog post
username	username(id) on PIXNET

Table 2. Attributes of writers

Attribute	Description
articles_count	number of the blog posts
blog body	HTML body of the blog
description	a short text to introduce the blog
display_name	nickname of the blogger
friends_count	number of the blogger's friends
has_ad	whether the blog has advertisement or not
hits_daily	daily number of hits
hits_total	total number of hits
hits_weekly	weekly number of hits
is_vip	total number of hits
keyword	keywords to introduce the blog
link	blog URL
name	username(id)
site_category	category of the blog
subscribers_count	number of subscribers to follow the blog

Table 3. Basic counts for reviews and writers

Instance	# of Instances	# of Paid Instances	Paid Ratio
Review	41,598	1,952	4.69%
Writer	2,054	245	11.93%

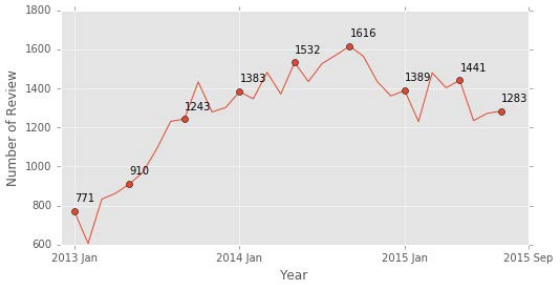

Figure 1. Number of reviews each month from January 2013 to August 2015

[2] https://developer.pixnet.pro
[3] https://www.crummy.com/software/BeautifulSoup

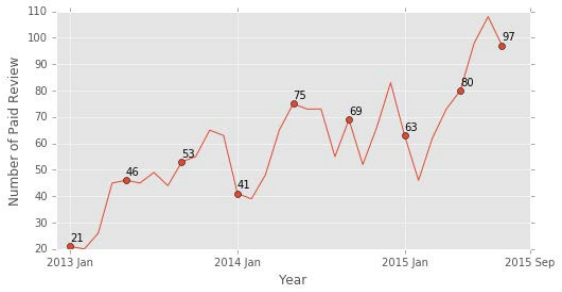

Figure 2. Number of paid reviews each month from January 2013 to August 2015

2.3 Comparison with the Other Datasets

To the best of our knowledge, this is the first research to study the opinion spam activity on blogs. People like to share their opinions on e-commerce or review sharing websites, whereas bloggers not only share opinions but also build their blogs to record memorable moments in their life or interact with their friends. We provide a comparison between this dataset and previous datasets from review websites or web forums as follows.

Rich and Longer Content: Most previous researchers only analyze the plain text while Pixnet users can apply various styles to text to emphasize some points using the rich text editor. The average length of an Amazon review is 123 words [5], and a blog post in our dataset has 824 words on average. A blog post usually has more paragraphs that include not only the dining experience but also authors' emotions.

Image: With the popularity of digital photography, bloggers put many photos in blog posts to enrich the content. Most reviews (41,432 of 41,598) in our dataset have at least one photo.

Bloggers Information: Bloggers provide personal/blog information like the brief description of the blogger/blog and the category of the blog while most of the review websites do not have such information.

Ground Truth: Ground truth is reliable because the authors label the ground truth themselves. We can assume the well-known writers who have written plenty of reviews label their paid review honestly since it will damage their reputation severely if they get caught cheating by other users. However, the ground truth may fail to detect dishonest reviews that are published by some throwaway accounts.

3 Data Exploration

We analyze the dataset to identify the sophisticated characteristics of paid reviews and paid writers. These writers are hired to promote food, beverage, and services for restaurants. Their behaviors differ from benign writers' in several ways.

3.1 Rich Review Content and Information

The goal of hiring paid writers is to introduce the restaurant in a positive and detailed view. We observed that paid reviews have longer content and more photos to tell readers the distinguishing features of the restaurant and describe their good dining experiences. We count the number of Chinese characters in each review and plot density and cumulative density distribution in Figure 3. The result shows that paid reviews have longer content. The similar result also holds for the number of images.

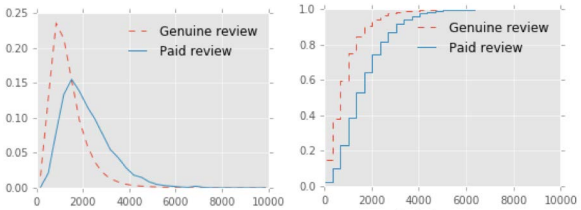

Figure 3. PDF and CDF of number of characters in reviews

Besides, we discovered that paid reviews have more information in titles. It includes the restaurant name, food, location, and a brief description. An informative title not only attracts readers' attention but also improves its ranking on search engines. Figure 4 shows that titles of paid reviews are longer.

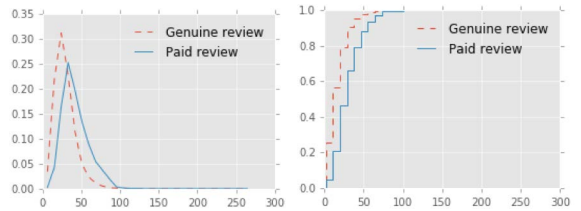

Figure 4. PDF and CDF of title length in characters

3.2 Writers' Popularity/Readership

Famous writers can grant more opportunities than unknown writers since the goal of restaurants is to influence potential customers. Figure 5 demonstrates that paid writers' blog has a larger number of hits. We can find similar results on the number of friends and subscribers. Paid writers have more friends and subscribers on Pixnet.

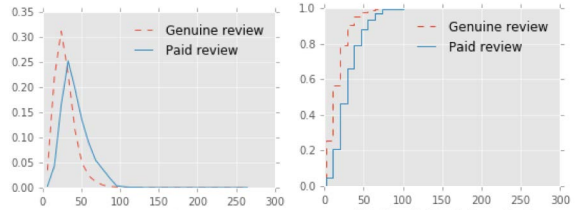

Figure 5. PDF and CDF of number of hits on blogs

3.3 Trust Establishment

Bloggers have to write a certain amount of blog post to increase their readership. A blog that has many blog posts and has run for

a period (e.g., several months or years) is more reliable for customers and more likely to get business attention. Figure 6 demonstrates that paid writers have more posts on their blog.

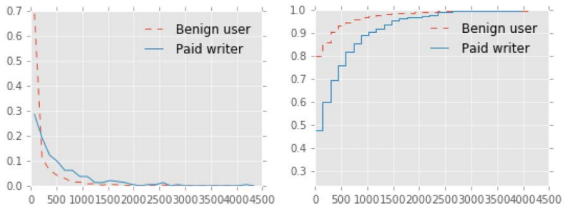

Figure 6. PDF and CDF of number of blog post on blogs

We then analyze the posting history of each paid writer. The posting history of a paid writer with n reviews can be viewed as a sequence $S = \{r_1, r_2, r_3 \ldots, r_{n-1}, r_n\}$, where the subscript denotes the posting order of a review. We calculate the *absolute position* and *relative position* for each review. If a review is the i-th item in the review sequence of a writer, its absolute position is i-1 and relative position is $(i-1)/n$. An example is shown as follows.

$S = \{r_1, r_2, r_3 \ldots, r_{n-1}, r_n\}$

$absolute\ position = \{0, 1, 2, \ldots, n-2, n-1\}$

$relative\ position = \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-2}{n}, \frac{n-1}{n}\right\}$

As shown in Figure 7, the distributions of relative locations are quite different between paid and genuine reviews. Paid reviews are more likely in the later part of a sequence. In other words, paid writers need to write genuine reviews to earn the trust first. For most paid writers, paid reviews only occupy a small proportion of their blog post (Figure 8) because too many paid reviews might arouse readers' disgust and decrease the willingness of restaurants to invite writers for writing reviews.

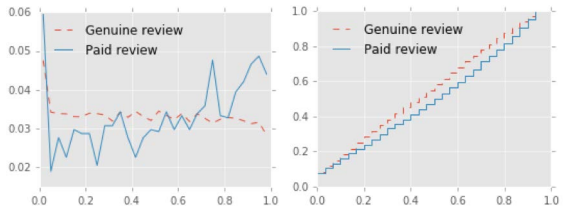

Figure 7. PDF (left) and CDF (right) of relative position in paid writer's posting sequence

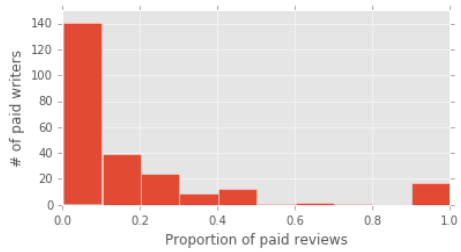

Figure 8. Number of paid writers with different proportion of paid reviews

3.4 Social Media Marketing

With the rise of social media, professional blogger has to embrace social media to interact with their readers. Table 4 reports the proportions of paid and genuine reviews, shared to the three social media platforms, Facebook, Plurk, and Twitter. It shows that Facebook is more popular than the other two social media platforms. Compared with genuine reviews, a larger proportion of paid reviews are shared to Facebook. We also investigate if the writers are using Facebook Fan Page by identifying the Facebook plugin in their blogs. The result in Table 5 confirms that paid writers are more likely to share blog posts on Facebook.

3.5 Pattern in Submission Time of Reviews

Chen and Chen [1] discovered that a higher percentage of spam posts are submitted during work time because spam activity is a job for opinion spammers. In contrast, paid writers have different patterns. Figure 9 shows the publishing time of paid reviews is similar to that of genuine reviews.

Table 4. Proportion of genuine and paid reviews shared to the three social media.

	Facebook	Plurk	Twitter
Genuine Review	8.59%	6.06%	0.62%
Paid Review	14.81%	5.84%	0.31%

Table 5. Statistics of Writers' Facebook usage.

	Reviews shared to Facebook at least once	Has Facebook Fan Page
Benign Writer	16.03%	4.48%
Paid Writer	28.16%	25.71%

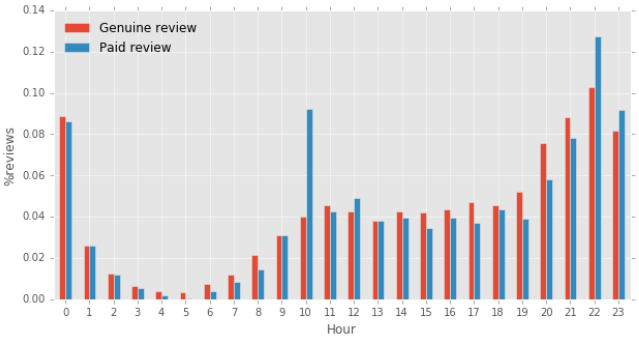

Figure 9. Proportion of reviews published throughout a day.

There are three explanations. Firstly, writing paid reviews is a part-time job for most paid writers. Secondly, in comparison with opinion spammers, paid writers work as individuals whereas some opinion spammers are involved in teamwork. Thirdly, publishing a blog post can be scheduled in the system. Paid writers may optimize the timing for a higher impact. The

publishing times of paid reviews concentrate on two peaks at 10 a.m. and midnight, which are peak hours of the Pixnet traffic. Posting at these time intervals can gain more attentions.

4 PAID REVIEW AND PAID WRITER DETECTION

Section 5.1 describes the method of paid review detection and shows the experimental results. The paid writer detection and the experimental results are shown in Section 5.2. We discuss several aspects including different types of features, learning techniques, experimental setup, evaluation metrics, and results.

4.1 Paid Review Detection

4.1.1 Proposed Features

As mentioned in Section 3.2, a blog post contains different types of data. We thus derive a variety of features from the blog post. In addition to the features suggested by previous work, we introduce new features, which are marked with †, based on the observations discussed in Section 4. In regard of the naming of features, the #_ prefix means *the number of*, and the %_ prefix means *the proportion of*. We scale each feature to zero mean and unit variance before feeding it to machine learning models.

(1) Text statistics: A set of features is derived from the main contents of reviews as introduced in Table 6, where those newly introduced features are marked with †.

(2) Metadata †: Table 7 shows the description of metadata features that estimate the quality and popularity of the review.

(3) Social Media Information †: As noted in Section 4.4, paid reviews are more likely to be shared to social media. We use three binary variables as features to indicate whether the review is shared to Facebook, Plurk, and Twitter.

(4) Image Information †: The occupational bloggers usually have a professional-grade camera and pay attention to photo quality. The resolution of their photos may be higher. In addition, we observe that some occupational bloggers host their photos on external image hostings other than on Pixnet for higher image quality. The detailed descriptions are shown in Table 8.

(5) Temporal Statistics †: To convert the publishing time into features, we use 24 + 7 = 31 binary indicators for each hour in a day and each day in a week. Relative and absolute positions proposed in Section 4.3 are also included.

(6) Content and Title †: We represent each review as a bag of words on a vector, and the weight of each word is its TF-IDF value. Similar procedure is applied to title.

(7) Tag and Category †: We extract all character bigrams in a tag or a category, and encode the tag/category by bag-of-bigrams with TF-IDF weighting.

(8) Word Embedding †: Mikolov et al. [12] proposed *word2vec*, a model generates word embedding for semantic modeling. We train a skip-gram model on our corpus and represent every review as a vector by calculating the average of the embedding vectors of each word in the review. The vector

Table 6. Description of text statistics features.

Feature	Description
#_all	Number of characters used in the review
#_digit	Number of digits characters
#_english	Number of English characters
#_line	Number of lines in the review
#_menu †	Number of how many times the review mention 菜單 (menu)
#_open †	Number of how many times the review mention 開幕 (opening)
#_punct	Number of punctuation characters
#_special	Number of non-alphanumeric characters
#_style	Number of font style used in the review
#_url	Number of URLs in the review
#_url_same_author †	Number of URLs from the same author
#_word	Number of words in the reviews
#_wspace	Number of white space characters
%_digit	Proportion of digits characters
%_english	Proportion of English characters
%_punct	Proportion of punctuation characters
%_special	Proportion of non-alphanumeric characters
%_wspace	Proportion of white space characters
title_length	Number of character of the title
has_facebook_plugin †	Whether Facebook Fan page plugin is used in the review
has_google_ad †	Whether Google AdSense code is used in the review

Table 7. Description of metadata features.

Feature	Description
has_address	Whether the author enters an address for the review
#_comment	Number of comments in the review
#_hit	Number of hits of the review
#_tag	Number of tags in the review
#_trackback	Number of other blog posts quote the review
cover_photo	Whether the review uses cover photo

Table 8. Description of image features.

Feature	Description
#_image	Number of images used in the review
max_width	Maximum width of images in the review
max_height	Maximum height of images in the review
external_host	Whether images in the review use external image hosting

size is set to 500, the window size is 5, the min-count is 300, the iteration is 20, and the number of negative examples is 20.

(9) LIWC † : Traditional Chinese LIWC constructed by Huang et al. [4] consists of 30 linguistic categories and 42 psychological categories. We count how many words appear in these 72 categories for each review and take the counting as *LIWC* feature.

4.1.2 Experimental Setup

Classifier: Logistics Regression (LR) and Support Vector Machine (SVM) classifiers are used.

Evaluation Metrics: A model can achieve high performance by predicting the majority class due to the imbalanced nature of our dataset. Thus, in paid review detection, F1 score on the paid review class is our main metric because we do not have any particular preference. Precision and recall are also reported.

Data Splitting: Writers are randomly divided into ten sets. For each set, we create a fold with all the reviews written by the bloggers in the set. Then we perform 10-fold cross validation to evaluate the performance. If paid reviews from the same author appear in both training set and test set, the model may learn the frequent words of the author instead of the real suspicious features. This setting prevents models from capturing user preference like writing habit.

Imbalanced Data: Our data is extremely imbalanced, so we adopt balanced weighting to improving model performance. We assigned different weights for misclassifying majority and minority class. Weights are inversely proportional to class sizes.

4.1.3 Results

We compare performances among all review features. Table 9 reports that the linear regression classifier with content features achieves a decent result and significantly outperform the random baseline (Precision = 0.046, Recall = 0.488, F1 = 0.084). The result also shows that the SVM with RBF kernel works well with Features (1)-(5). Linear SVM (All) achieves the best F1 score 0.549.

Table 9. Experimental results using LR, Linear SVM, and RBF SVM with different review features.

Features	LR			Linear SVM			RBF SVM		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Features (1)-(5)	0.139	0.746	0.234	0.137	0.723	0.231	0.157	0.669	0.254
Content	0.410	0.640	0.500	0.430	0.588	0.497	0.032	0.388	0.059
Title	0.202	0.651	0.308	0.180	0.575	0.274	0.021	0.083	0.033
Tag	0.149	0.500	0.230	0.127	0.412	0.194	0.040	0.669	0.075
Category	0.099	0.293	0.148	0.095	0.291	0.143	0.029	0.291	0.052
LIWC	0.138	0.721	0.232	0.169	0.632	0.266	0.136	0.703	0.228
Embedding	0.209	0.823	0.334	0.208	0.824	0.332	0.056	0.242	0.091
All	0.452	0.629	0.526	0.527	0.573	0.549	0.102	0.699	0.178
All but removing honest keywords	0.464	0.613	0.529	0.509	0.559	0.533	0.101	0.720	0.177

Content features provide precise prediction. There may exist some frequently used words or writing habit in genuine reviews. Thus, we decide to find which words play the important role in distinguishing between paid and genuine reviews. We train a linear SVM model on content TF-IDF and examine the importance of each word by looking at its coefficient. The left word cloud in Figure 10 contains words with the lowest coefficients, which are the strongest genuine review indicators. The font size of each word positively correlates to the absolute value of its coefficient. The right word cloud contains the suspicious words, which are the strongest paid review indicators. The top suspicious words are frequent words from honest paid reviews, which authors point out that they are invited by the restaurant to write reviews. We find that the most genuine words represent the usual dining behavior that paid writers are not likely to do. For example, paid writers do not need to pay for the meal, so all words related to payment only appear in the most genuine words. Similarly, the words related to reservation only appear in the most genuine words since paid writers, who are invited by restaurants, do not have to make the reservation. Moreover, it is worth noting that non-positive adjectives are more likely to appear in genuine reviews.

Figure 10. Most "genuine" (left) and "paid" (right) words.

We remove the honest keywords and conduct detection experiments again. The last row in Table 9 shows all the metrics drop slightly when the honest keywords are removed, and the performance is still decent.

4.2 Paid Writer Detection

The adopted features are listed below. Some of them are similar to the features for paid review detection. The new features we introduced for blogging are also marked with † .

Blog Information † : Table 10 lists blog information features.

Social Media Marketing † : We estimate social media usage by the proportion of reviews shared to Facebook, Plurk, and Twitter.

Content Similarity: We use the bag-of-words with TF-IDF weighting to represent reviews of each blogger, and calculate the cosine similarities between reviews. The maximum of the similarities is taken as the Content Similarity feature.

Paid Indicator: The Paid Indicator feature is computed by taking the maximum of the paid review probability of all reviews published by the writer.

Description: For each blog description, we first segment the text and transform it into a vector encoded in the bag-of-words with TF-IDF weighting.

Keyword: We take character bigrams of keywords for each writer and represent keywords in a vector of bag-of-bigrams with TF-IDF weighting.

The classifiers LR, SVM with linear kernel, and SVM with RBF kernel are employed in paid writer detection experiments. As paid review detection, paid writer detection also has the imbalanced problem. All models use balanced weighting to handle the imbalanced issue. We have split users into ten sets in paid review detection. The same ten sets are used to perform 10-fold cross validation in paid writer detection. F1 score on paid writer is the main metric for evaluation. Precision and recall are also reported.

Table 10. Description of blog information features.

Feature	Description
#_friend	Number of friends the blogger has
#_subscriber	Number of subscribers the blogger has
#_hit	Number of hits on the blog
#_blog_post	Number of posts on the blog
#_review	Number of reviews in the dataset
has_ad	Whether the blogger has native advertisement from Pixnet
is_vip	Whether the blogger pays for VIP plan to use extra services
has_external_ad	Whether the blogger put Google AdSense Advertisement in the blog
has_email	Whether the blogger introduction contains email addresses
first_review_time	Submission time of the first review
description_length	Number of characters of the blogger description
#_keyword	Number of keywords of the blog

Table 11. Experimental results using LR, Linear SVM, and RBF SVM with different writer features.

Features	LR			Linear SVM			RBF SVM		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Blog Information	0.344	0.743	0.470	0.348	0.751	0.475	0.319	0.808	0.457
Social Media Marketing	0.277	0.273	0.275	0.287	0.269	0.278	0.244	0.384	0.298
Paid Indicator	0.473	0.824	0.601	0.488	0.812	0.609	0.446	0.857	0.587
Content Similarity	0.329	0.710	0.450	0.337	0.698	0.454	0.295	0.780	0.428
Description	0.248	0.416	0.311	0.247	0.420	0.311	0.118	0.486	0.190
Keyword	0.233	0.514	0.321	0.234	0.518	0.322	0.118	0.486	0.190
All	0.487	0.767	0.596	0.496	0.771	0.604	0.467	0.780	0.584
All but removing paid indicator	0.400	0.665	0.499	0.407	0.657	0.502	0.385	0.682	0.492

Table 11 shows the results by using different features. Performances of three learning models are close, and Paid Indicator is the strongest feature. The performance of using paid indicator alone is slightly higher than the performance of using all the features. It demonstrates that the paid review detection model can help spot paid writers. Other features work well in some sense. Using Blog Information and Content Similarity significantly outperforms the random baseline (Precision = 0.120, Recall = 0.501, and F1 = 0.194). That supports our observations.

In this study, we also explore the collective detection by using typed Markov Random Fields (T-MRF) [9]. T-MRF aims to integrate paid review detection with paid writer detection by leveraging the relational data. However, experimental results show that the collective detection model does not get better performance in this dataset.

5 CONCLUSION AND FUTURE WORK

In this paper, we build a novel dataset for paid reviews and paid writers detection. We find that advertisers hire bloggers with a large readership to promote their targets. Paid writers are more popular than benign users and tend to produce high-quality contents to promote the target restaurants. Besides, they need to write a certain amount of genuine reviews to earn readers' trust. With the rise of social media, paid writers also use the social network to market themselves.

We conduct paid review and paid writer detection using supervised learning techniques. In paid review detection, we propose a set of features from contents and metadata. The content-based features achieve the best performance, and other proposed features also significantly outperform the baseline model. In paid writer detection, we put the efforts to capture writer' behavior. Our results show that the features Paid

Indicator and Content Similarity work well. That supports our observations.

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-104-2221-E-002-061-MY3 and MOST-105-2221-E-002-154-MY3.

REFERENCES

[1] Chen, Y. R. and Chen, H. H. 2015. Opinion Spam Detection in Web Forum: A Real Case Study, In *Proceedings of the 24th International Conference on World Wide Web*, ACM, 173-183.

[2] Chen, Y. R. and Chen, H. H. 2015. Opinion Spammer Detection in Web Forum. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 759-762.

[3] Harris, C. G. 2012. Detecting Deceptive Opinion Spam Using Human Computation. *Human Computation AAAI Technical Report WS-12-08*, Association for the Advancement of Artificial Intelligence, 87-93.

[4] Huang, C. L., Chung, C. K., Hui, N., Lin, Y. C., Seih, Y. T., Lam, B. C. P., Chen, W. C., Bond, M. H., and Pennebaker, J. W. 2012. The Development of the Chinese Linguistic Inquiry and Word Count Dictionary. *Chinese Journal of Psychology* (2012), 185-201.

[5] Jindal, N. and Liu, B. 2007. Review Spam Detection. In *Proceedings of the 16th International Conference on World Wide Web*, ACM, 1189-1190.

[6] Jindal, N. and Liu, B. 2008. Opinion Spam and Analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ACM, 219-230.

[7] Ko, M. C. and Chen, H. H. 2016. Analysis of Cyber Army's Behaviours on Web Forum for Elect Campaign. *Information Retrieval Technology*, Springer, Vol 9460, 394-399.

[8] Lai, C.L., Xu, K.Q., Lau, R. Y. K., and Li, Y. 2010. Toward A Language Modeling Approach for Consumer Review Spam Detection. In *Proceedings of IEEE International Conference on E-Business Engineering*, 1-8.

[9] Li, H., Mukherjee, A., Liu, B., Kornfield, R., and Emery, S. 2014. Detecting Campaign Promoters on Twitter using Markov Random Fields. In *Proceedings of the 2014 IEEE International Conference on Data Mining*.

[10] Li, H., Chen, Z., Mukherjee, A., Liu, B., and Shao, J. 2015. Analyzing and Detecting Opinion Spam on a Large-scale Dataset via Temporal and Spatial Patterns. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 634-637.

[11] Lim, E. P., Nguyen, V. A., Jindal, N., Liu, B., and Lauw, H. W. 2010. Detecting Product Review Spammers using Rating Behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ACM, 939-948.

[12] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality, In *Advances in Neural Information Processing Systems*, 3111-3119.

[13] Mukherjee, A., Liu, B., Wang, J., Glance, N., and Jindal, N. 2011. Detecting Group Review Spam. In *Proceedings of the 20th International Conference Companion on World Wide Web*, ACM, 93-94.

[14] Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. 2013. What Yelp Fake Review Filter Might Be Doing? In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 409-418.

[15] Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 209-319.

[16] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. 2007. The Development and Psychometric Properties of LIWC2007. Austin, TX, LIWC.Net.

[17] Tang, D., Qin, B., Liu, T. 2015. Learning Semantic Representations of Users and Products for Document Level Sentiment Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, 1014-1023.

[18] KC, S. and Mukherjee, A. 2016. On the Temporal Dynamics of Opinion Spamming: Case Studies on Yelp. In *Proceedings of the 25th International Conference on World Wide Web*, 369-379.