Characterizing Users in an Online Classified Ad Network

Muhammad Waqar
Department of Computing Science
University of Alberta
mwaqar@ualberta.ca

Davood Rafiei
Department of Computing Science
University of Alberta
drafiei@ualberta.ca

ABSTRACT

Unlike online social networking sites (e.g. Twitter and Facebook) which are heavily used for disseminating content and sharing information between users and shopping sites (e.g. Ebay) where buyers and sellers are reviewed, the flow of information between users such as buyers and sellers in a classified ad network is very limited. Characterizing users or assigning them to some classes in one such network is challenging due to the sparsity of the data about users, the vague separation of user classes and sometimes the tendency of users to hide or misrepresent their profile information.

In this paper, we study the information revealed in the ads posted to an online classified ads site and analyze the behaviour of users posting those ads; our study is conducted using data collected from Kijiji over a year. We study the problem in the context of one specific task where we seek to detect if a user posting an ad belongs to one of the two classes *business* and *non-business*, based on the ads the user has posted. We study an approach based on user profiling, where given statistics on how an ad mentions terms and features from a class profile, the affinity of an ad (and subsequently a user) to a particular class is determined. We report the effectiveness of this approach in detecting user classes solely based on the information revealed in their ads and study the impact of the profile size on the accuracy. In the absence of labeled training data, we show that a simple bootstrapping technique with only a few n-grams as a seed set can give nearly good results in terms of F-measure. We further report our experiments on characterizing the collective behavior of users in posting ads and some of the distinctive usage patterns that emerge.

Categories and Subject Descriptors

H.2.8 [**Database management**]: Database applications–Data mining.

WIMS '16 June 13-15, 2016 Nîmes, France

DOI:

Keywords

classified ads, user modeling, user profiling, social networks.

1. INTRODUCTION

User 64073399 has posted a classified sales ad on Kijiji [1] for a 2008 Honda Accord, and the ad claims to be "for sale by owner." How much can this be trusted? Is there an independent way of verifying this claim? The tremendous growth of the World Wide Web (WWW) combined with lower costs and greater convenience have been some of the factors driving individuals away from traditional print ads and more towards online classified advertising. Although these ads are mostly placed by private individuals, many businesses are also using this medium for the promotion of their products and/or services, finding the right job applicants, etc.

Ascertaining if a user in a classified ad network is a business or a private individual involves various challenges. First of all, the distinction between the two classes of users is often vague and so is some of their postings. Many users who appear to run a business using the network do not explicitly state this fact. Additionally, the distribution of users in the two classes is highly imbalanced, since such networks are mostly geared towards individuals than businesses who can avail many other forms of advertising too. Moreover, the data posted by the users in many cases is extremely sparse, as most of the users do not use the network on a regular basis, but only when a specific need surfaces.

Despite these challenges, such a separation of users can have many interesting applications. For example, in a system that traditionally involves no user feedback, it gives the users better information about the nature of a seller. The government may use the data to better regulate certain sectors of business which may not be reachable otherwise. An online ad network provider may use this information to analyze its different pricing strategies. Yet another application is populating Web directories which can take considerable time and effort if done manually. Such data about users can also help in better understanding the dynamics of the classified ad networks.

In this paper, we study the problem of characterizing users based on their ads; the more concrete problem is detecting if a user belongs to one of two given classes. A *user* here refers to someone who posts an ad and often represents sellers; but a buyer may also post a "wanted" ad and be considered as a user. This problem can be cast as a classification task where given a user and his posted ads, the goal is to detect if the

[1] http://www.kijiji.ca

user is a *business* or *non-business*. Despite the large body of work on user modeling and classification (as some listed in the next section), to the best of our knowledge, this is the first such study on classified ad postings. A caveat is that it is possible for business users to use the ad network for their personal use as well, and classifying these users can be subjective. When a single label is sought, such users are treated as business in this paper. Our contributions can be summarized as follows: (1) we study the problem of profiling users in an online classified ad network; (2) we present a method for detecting business users based on their ads; (3) we evaluate our methods using a crawl of a real network and report some of our findings and analysis as well comparisons to a few competitive baselines; (4) we present a bootstrapping method to collect large volumes of training data; (5) finally we show how some of the distinctive usage patterns of a desired user class can be detected.

Organization The rest of this paper is organized as follows. We review the related work in Section 2 and present our experimental setup in Section 3. Our methodology for classifying users based on their ads content is presented and evaluated in Section 4, and the posting behavior of the users is analyzed in the same context in Section 5. The paper is summarized and concluded in Section 6.

2. RELATED WORK

Our work relates to the areas of *text classification*, *user modeling in social media* and *social network analysis*.

Text classification Since our work utilizes the content of an ad to determine its affinity to *business* and *non-business* classes, the large body of work on text classification is relevant. Early work in this field focused on categorizing documents by topics, and a comprehensive survey of such techniques can be found in [Aas and Eikvil, 1999, Berry and Castellanos, 2004]. Examples of topical text classification can be found in classifying news stories and blog posts (e.g. [Sun et al., 2007]). There has been also some recent interest in the field of non-topical classification. A closely related work to ours is that of Makazhanov et al. [2014] which uses the interactions with a party to determine the political preference of Twitter users. However, unlike the work of Makazhanov et al., we do not classify the users directly, but aggregate the results of our ad classification, that uses a similar technique, to predict a label for each user.

User modeling in social media Our work builds models of users in an online classified ad network based on not only the content of their ads but also their posting patterns, hence it is related to similar modeling exercises in social media. Liu et al. [2010] use a Bayesian model to predict users' news interests based on their past activities on the web as well as the current news trends; they later utilize these preferences for personalized news recommendation. Abel et al. [2011] study the same problem in the context of the Twitter network, utilizing tweets posted by users to infer their preferences. Carmagnola et al. [2007] argue that leveraging tags can be helpful for systems to learn more about their users. Schöfegger et al. [2012] analyzed this tagging behavior of users in a social academic network to predict their research discipline. Stoyanovich et al. [2008] propose a model to infer users' interests by leveraging the tags generated by not only the users themselves but also their social friends. These studies are closely related to the extensive body of work on recommender systems, which learn a model of users' interest based on their past behavior. A relatively comprehensive survey of such works, covering the work before 2005, can be found in [Adomavicius and Tuzhilin, 2005]. More recently, strategies have been put forward for modeling users from heterogenous sources on social web to improve the quality of recommender systems ([Abel et al., 2013]).

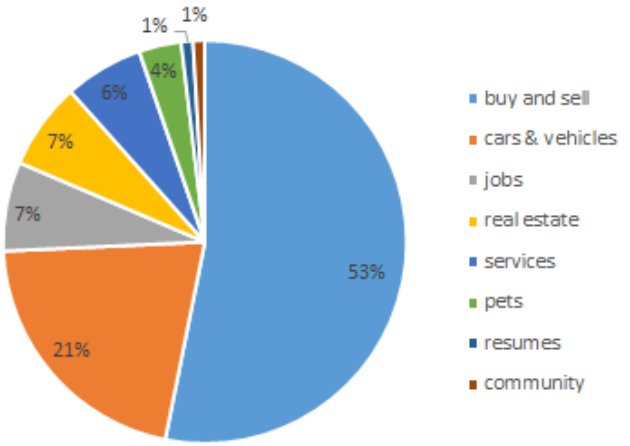


Figure 1: Distribution of the ads in various categories

Extensive work has been done on leveraging user models to find other users with similar behavior. Benevenuto et al. [2010] collect a vast Twitter dataset and identify a number of features related to tweet content and user behavior which are then used to detect spammers. Similar approaches are also used to identify hidden paid posters in online communities [Chen et al., 2013] and spammers as well as content promoters in online video social networks [Benevenuto et al., 2009].

Social network analysis Online classified ads also exhibit some of the traits of a social network in the way users interact but perhaps implicitly, for example by posting similar ads (listing the same or similar items), tagging the same location for the ads, responding to other users' ads, etc. Therefore, the body of work on social network analysis is loosely relevant. Related work includes the literature on finding groups of users or *communities* whose members share a similar profile, i.e., exhibit a similar behavior in their interactions [Orwant, 1994]. The proposed techniques include clustering [Fisher, 1987], modularity based approaches [Clauset et al., 2004], graph based partitioning [Ng et al., 2002], clique percolation [Palla et al., 2005], etc. Some of these techniques may also be applicable in the setting of a classified ad network, to find user groupings with similar interests.

3. DATA COLLECTION AND EXPERIMENTAL SETUP

Data collection For all the experiments in this work, we collected advertisements from Kijiji, a popular online classified ad service that allows users to post free classified ads in different categories. We chose Kijiji for our work because of its popularity in Canada, its wide presence in multiple other countries and the fact that it allows users to register an account with the website. This allows us to connect each ad to the user who posted it and view other ads posted by

the same user, hence offering all sorts of benefits in modeling users and tracing their activities. Our work can as well be adapted to any other classified ad network such as Craigslist as long as users posting the ads can be identified.

A crawler was built to collect ads from Edmonton Kijiji [2]; the crawler was run once every day over a period of 9 months starting May 1, 2013. During the crawling period, all posted ads that were active at the time of the crawl were fetched and various fields including ad id, title, category (stripped of location information), listing date, description (in plain text) and user id were extracted. The dataset had some very diverse categories, such as *buy and sell* which included buy and sell ads ranging from books to an entire business, whereas others were relatively limited in their scope, such as *pets*. Our experiments utilize *buy and sell* from the diverse categories and *cars & vehicles* from more specialized categories; these are the two largest categories in terms of the number of ads, as shown in Figure 1. Also *buy and sell* presents the highest nature of imbalance in terms of *business* and *non-business* classes, making the user classification task most challenging.

Ground truth A random sample of 5,000 users were selected from the dataset for manual labeling. After examining the postings of users, the annotators marked 157 users (3.14%) as *business* and 4,634 users (92.68%) as *non-business*. There were 209 users (4.18%) where the annotators were not sure about, and these users were marked as *unknown* and were ignored in our experiments.

Dealing with imbalanced data We follow the recommendation by Klement et al. [2009] and combine random undersampling with an ensemble of classifiers. Each classifier in the ensemble is trained on a balanced sample of training set which is obtained by randomly under-sampling the majority class (*non-business*) while preserving the complete minority class (*business*). The individual classifiers are combined by averaging their predicted confidence.

4. FROM AD CONTENT TO USER CLASSIFICATION

Our hypothesis in classifying users is that the ads posted by business users are likely to reveal traits that are different from those of non-business users. In light of this hypothesis, the problem of user classification involves "detecting for a given ad, if the ad has more affinity for *business* class or *non-business* class." Treating each label as an abstract concept, we associate with each class a ranked list of weighted terms, referred to as the *class profile*. Given an ad that *mentions* terms from a profile, we seek to characterize the ad as either business oriented or non-business oriented.

4.1 Building Profiles

Let l denote the class label of an ad that we want to predict. In our case, $l \in \{business, non\text{-}business\} = L$. Let D_l denote the collection of all ads with class label l. We refer to this as the class corpus. The entire corpus, therefore, is denoted by $D = \{D_l \mid \forall l \in L\}$ and its vocabulary is denoted as V. We build a language model (LM) for each class as well as one for the entire corpus, and calculate the KL-divergence between the LM probabilities of each class corpus and the entire corpus, as done in some early work as well (e.g. [Shmueli-Scheuer et al., 2010]). The divergence score of an individual term can then be used as a measure of importance of the term to a specific class.

tf-idf scores are used to calculate term probabilities for a corpus, and the marginal probability of a term is calculated and normalized as:

$$P(t \mid D) = \overline{tf}(t, D) udf(t, D),$$
$$P^N(t \mid D) = \frac{P(t \mid D)}{\sum_{t \in V} P(t \mid D)}$$

where $\overline{tf}(t, D)$ represents the average frequency of term t in documents (ads) in D and $udf(t, D) = df(t, D)/|D|$. $df(t, D)$ denotes the document frequency of t in D.

For class corpora, initial term weights are calculated and normalized as:

$$w(t \mid D_l) = \overline{tf}(t, D_l) udf(t, D_l) idf(t, D),$$
$$w^N(t \mid D_l) = \frac{w(t \mid D_l)}{\sum_{t \in V} w(t \mid D_l)}$$

where $idf(t, D) = \frac{|D|}{1+df(t,D)}$ is the inverse document frequency of t in D.

To account for missing terms in a class corpus, the weights are smoothed as:

$$w^S(t \mid D_l) = (1 - \lambda) w^N(t \mid D_l) + \lambda P^N(t \mid D)$$

where the normalization factor λ is set to 0.001.

Finally, the probability of a term in the LM of a class corpus is

$$P(t \mid D_l) = \frac{w^S(t \mid D_l)}{\sum_{t \in V} w^S(t \mid D_l)}.$$

Now the KL-divergence between probability distributions of corpus LM and class LM can be calculated as:

$$KL_p(P(t \mid D_l) \| P(t \mid D)) = \sum_{t \in V} P(t \mid D_l) \ln \frac{P(t \mid D_l)}{P(t \mid D)}.$$

Instead of the entire content difference, as represented by the sum in the above equation, we are more interested in the divergence between corpus LM and class LM for each term. This importance score for a term is:

$$I(t, l) = P(t \mid D_l) \ln \frac{P(t \mid D_l)}{P(t \mid D)}.$$

The higher the importance score of a term is, the more it will deviate from the common vocabulary and be more important to a particular class. Accordingly, we select the top-N terms as the class profile.

For the *buy and sell* category in our dataset, the top three business terms were "we have," "for more" and "http www" and the top three non-business terms were "posted with," "kijiji mobile" and "i have." The business terms for *cars & vehicles* were "see more," "our dealership" and "serve you" and the non-business terms were "posted with," "kijiji mobile" and "selling my."

4.2 Methodology

We employ a one-vs-all classification strategy with a binary classifier trained for each class label; each ad is assigned the class with the highest predicted confidence. Ads that do not mention any terms from a class profile or have a predicted confidence score of less than 0.5 are termed as

[2]http://edmonton.kijiji.ca

unknown and are ignored for that particular class in the ads classification task.

A few features are used to describe the relationship between ads and class profiles. One feature is the number of mentions, chosen based on the idea that the more an ad mentions terms from a class profile, the more likely it is tilted towards that class. Other features include the average weight of mentions and the average, the minimum and the maximum rank of the mentions; the higher a term is ranked in a class profile, the more relevant and distinctive it is to that class. An ad can mention terms from both business and non-business classes, hence a relative weighting scheme is used [Makazhanov et al., 2014].

Training data was collected for each top-level category; we expected the same or similar class terms for same class of ads within a top-level category (e.g. *cars & vehicles*) but possibly different for ads from different categories (e.g. *cars & vehicles* and *buy and sell*). In a random sample of 1,858 ads from the *buy and sell* category, 150 (8.07%) were deemed business, 1,585 (85.31%) were deemed non-business and the remaining 123 (6.62%) were marked as unknown by our annotators. Similarly, in a random sample of 756 ads from the *cars & vehicles* category, 150 (19.84%) were deemed business, 578 (76.45%) were deemed non-business and the remaining 28 (3.7%) were labeled unknown.

4.3 Ad Classification results

In a 10-fold cross-validation experiment, we evaluated the performance of our ad classification, in terms of Precision (P), Recall (R) and F-measure (F). We experimented with three classifiers: decision tree based Random Forest (RF), SVM based SMO and Logistic Regression (LR). The parameters were set to their default values in Weka. We used two approaches for training our classifiers: (1) random undersampling with an ensemble of classifiers (RUSEC) detailed in Section 3 and (2) using the imbalanced data (IMB) as is. As our vocabulary, we used unigrams and bigrams from ad titles and ad descriptions with a profile size set to 100 in our experiments. While building class profiles, we deliberately ignore the numbers and rare terms (those occurring in two or less ads).

As shown in Table 1, the classifiers trained using the RUSEC approach have a much higher recall for business class as compared to the ones trained using IMB. On the contrary, IMB classifiers achieve a higher recall for non-business class than the RUSEC ones. This trend is to be expected. Since the training data in the IMB approach is imbalanced, the respective classifiers optimize predictions for the dominant class. This is not the case with the RUSEC training method since the data given to the classifier for training is balanced.

However, a reversed trend is noticed for the precision; RUSEC classifiers have a much lower precision for the business class as compared to the IMB and vice versa. This behavior follows from the argument mentioned earlier. Since IMB classifiers are optimized to cater for the dominant class, they tend to predict an instance as belonging to the minority class when there is an overwhelming evidence for this action. On the contrary, RUSEC classifiers tend to over-represent the minority class in their final predictions in comparison to its true underlying distribution.

Moreover, it can be seen that the recall of the non-business class is much lower for *buy and sell* compared to *cars & vehicles* especially for the RUSEC approach. We believe this is because of the fact that *buy and sell* is a much more diverse category, and the probability that a few top n-grams of the profile can capture a sufficient vocabulary for non-business users in this category is lower compared to *cars & vehicles*.

Finally, we notice that the precision of the business class is much lower for *buy and sell* than that for *cars & vehicles*. This is due to the fact that the dataset for the *buy and sell* category is much more imbalanced than that for the *cars & vehicles*. Therefore, a similar percentage of non-business misclassifications has a stronger impact on the precision of business users for *buy and sell* than *cars & vehicles*.

Impact of profile size To observe the effect of using a larger profile on the classification performance, we selected the best performing classifiers for both training approaches (i.e. Random Forest for RUSEC and Logistic Regression for IMB) and repeated the same experiment by increasing the profile size from 100 to 1,000 with an increment size set to 100.

Figure 2 shows the F-measures for business and non-business classes for both categories and training techniques. In all the cases, we found that having a larger profile does not impact the results significantly. Usually a subtle improvement is noticed in the F-measure in the first few iterations (the only exception being IMB training for *buy and sell* where it increases by nearly 8% initially). However, as we continue to increase the profile size, scores become stable and even start decreasing. This trend is not surprising, since the lower the terms are in a profile, the more they are a part of common users vocabulary rather than being distinctive for a particular class.

Feature analysis It can be observed from Figures 3a and 3b that while the ads that mention terms from a class profile excessively (≥ 80) are almost exclusively tilted towards that class, the overlap among positive and negative examples increases when the *number of mentions* (referred to as T feature domain) is low. However, using *relative number of mentions* (referred to as R domain), the instances can be differentiated very easily, making it the most effective feature for both categories.

Likewise, ads that mention top-ranked terms from the profile of a particular class usually indicate a stronger relevance towards the respective class. Accordingly, we find that *relative minimum rank of mentions* is an important feature for both categories.

4.4 Classifying users

Since we adopted a one-vs-all classification strategy to classify the ads, the classifiers output not only a prediction for an ad but also the confidence of the prediction i.e. the probability of the ad actually belonging to that class. We treat these probability values as noisy samples of a 'true' value, which is estimated as the arithmetic mean, i.e.

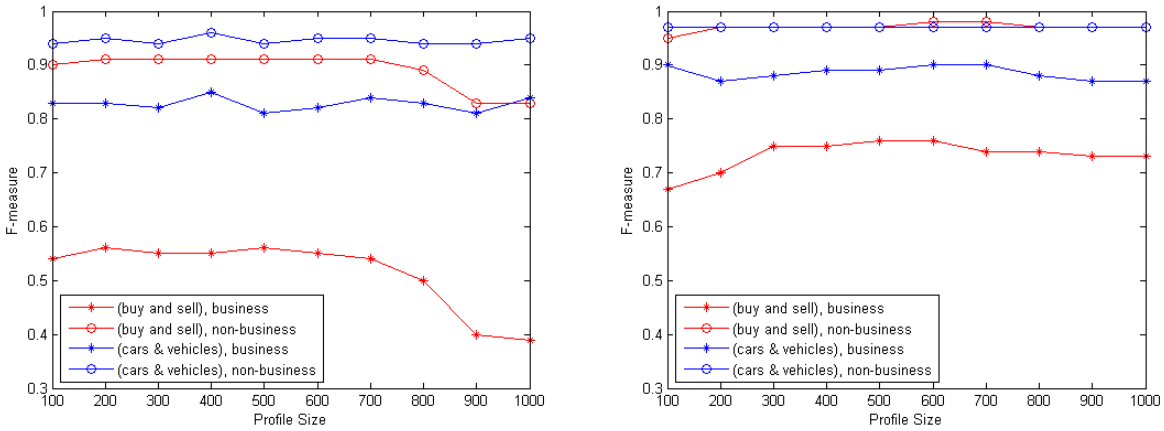
$$Confidence(u, c) = \frac{\sum_{x \in a_{u,c}} Confidence(x, c)}{|a_{u,c}|}$$

where $a_{u,c}$ represents the ads of the user u that mention terms from the profile of class c. A user u is assigned the class label with the highest confidence value.

In our evaluation, we trained the classifiers using both RUSEC and IMB approaches with the profile size fixed at 200 and classified user ads in our dataset. We used the

		Business			Non-Business			Business			Non-Business		
		P	R	F	P	R	F	P	R	F	P	R	F
		buy and sell						*cars & vehicles*					
RUSEC	LR	0.38	0.87	0.53	0.99	0.81	0.89	0.77	0.91	0.83	0.98	0.9	0.94
	SMO	0.35	0.89	0.5	0.99	0.8	0.89	0.82	0.87	**0.84**	0.97	0.94	**0.96**
	RF	0.41	0.89	**0.56**	0.99	0.82	**0.9**	0.8	0.89	0.84	0.97	0.92	0.95
IMB	LR	0.84	0.56	**0.67**	0.96	0.94	**0.95**	0.96	0.85	**0.9**	0.97	0.97	**0.97**
	SMO	0.84	0.45	0.58	0.95	0.95	0.95	0.94	0.79	0.86	0.96	0.97	0.97
	RF	0.81	0.47	0.59	0.95	0.94	0.95	0.97	0.81	0.88	0.96	0.98	0.97

Table 1: Results of the ad classification with the profile size is set to 100



(a) RUSEC (b) IMB

Figure 2: Ad classification performance varying the profile size

following baselines for a comparison:

Dealer (DE) Kijiji allows users to self-identify themselves as businesses through ad attributes in certain categories. In case of the two largest categories, this attribute is present in the *cars & vehicles* and *buy and sell/furniture* categories only. Note that for this baseline, we consider such attributes in "offering" ads only i.e. ads in which an item or service is offered.

Short URL (SU) Business users usually post a link to their official website in ad description to give viewers more information about their services. On the other hand, links present in the ads of non-business users are mostly to pages on the manufacturer's official website detailing the features of the item being sold; as such, these pages are often buried deep down the website's primary address. We treat the URLs that contain no directory paths after the main address (network location) as "short URLs". Thus http://webdocs.cs.ualberta.ca/ is a short URL whereas http://webdocs.cs.ualberta.ca/~mwaqar/ is not. By this baseline, a user is classified as a business if any of his ads contain a "short URL" and non-business otherwise.

One ad per week (1AD) We expect business users to use the network *frequently* to promote their enterprise as opposed to non-business users who would be anticipated to post on the network only when a need arises. According to this baseline, we define *frequently* as having posted at least one ad per week. Thus, a user is classified as business if his postings match this criteria and non-business otherwise.

The result of the user classification task is presented in Table 2. As discussed in Section 4.3, RUSEC-trained classifiers are able to detect more businesses correctly (including the baselines) while those trained with the IMB strategy achieve a higher precision for business users (among our classifiers). RUSEC classifiers are not able to make up for what they lose in precision for business class with even the highest recalls and are dominated by the IMB trained classifiers in terms of F-measure. Moreover, in either training approach, Random Forest gives the most impressive results. Additionally, all the baselines achieve low business recall, however, *Dealer* and *Short URL* have an impressive precision, even higher than our classifiers in case of the former. Finally, all our classifiers have a higher F-measure for business class than the baselines.

In order to ascertain if there is a significant improvement in terms of F-measure using our method, we applied the paired t-test on results obtained from each classifier and each baseline method. The null hypothesis was: *our method has no significant improvement*. According to t-test results, we obtain $p < 0.001$ for all combinations of our classifiers and the baselines, meaning that the difference in performance is statistically significant.

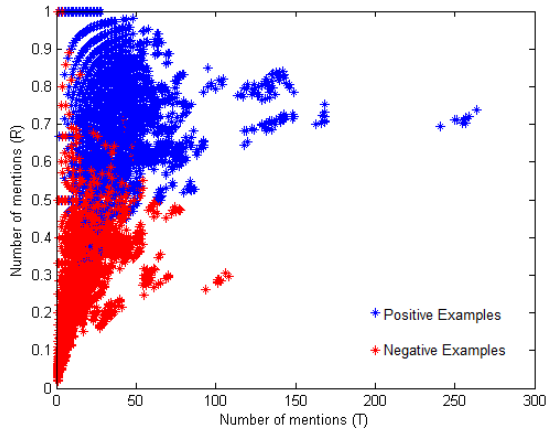
4.5 Using Unlabeled Data

Let us consider the scenario where labeled ads data is unavailable, hence the supervised classification scheme presented earlier is not possible. This scenario is not very far-fetched as significant time and effort is required to collect the training data by hand.

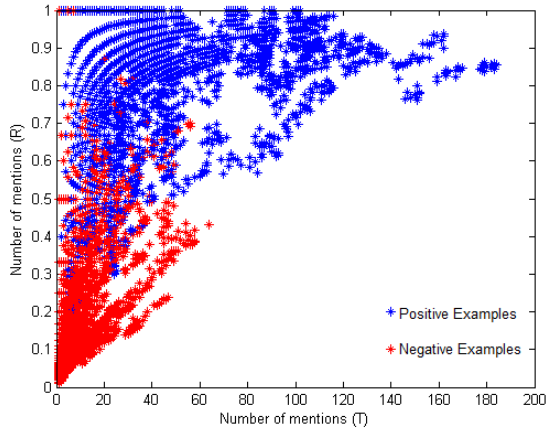
To tackle this challenge, we employ a simple bootstrapping heuristic where we provide a few n-grams to act as

		Business			Non-Business		
		P	R	F	P	R	F
RUSEC	LR	0.2	0.59	0.3	0.98	0.92	0.95
	SMO	0.21	0.58	0.3	0.98	0.92	0.95
	RF	0.24	0.57	**0.34**	0.98	0.94	**0.96**
IMB	LR	0.53	0.39	0.45	0.98	0.99	0.98
	SMO	0.67	0.32	0.44	0.98	0.99	0.99
	RF	0.73	0.36	**0.48**	0.98	1.00	**0.99**
Baselines	DE	0.93	0.17	0.28	0.97	1.00	0.99
	SU	0.68	0.16	0.26	0.97	1.00	0.99
	1AD	0.04	0.27	0.08	0.97	0.8	0.88

Table 2: User classification performance results with the profile size set to 200



(a) Number of mentions (T) / Number of mentions (R) for *buy and sell*



(b) Number of mentions (T) / Number of mentions (R) for *cars & vehicles*

Figure 3: Feature break-down into domains

		Business			Non-Business		
		P	R	F	P	R	F
RUSEC	LR	0.2	0.58	0.29	0.99	0.92	0.95
	SMO	0.18	0.59	0.27	0.99	0.91	0.95
	RF	0.26	0.51	**0.35**	0.98	0.95	**0.97**

Table 3: User classification performance results using unlabeled data and the profile size set to 200

a seed set with the expectation that these n-grams will be prevalent in business oriented ads. Accordingly, all the ads in the dataset that contain any n-gram from the seed list are treated as business ads. We thus obtain a labeled dataset to act as the training data for the classification of user ads and the normal ad classification methodology follows.

For our evaluation, we used the classifiers trained using the RUSEC approach with the profile size set to 200. We used only 4 n-grams as a seed set: *satisfaction*, *guaranteed*, *priority* and *hours of operation*. We also set a limit on the maximum number of ads containing any n-gram from the seed list that can be selected per user. This step was taken to prevent the language model of a class from becoming biased towards only a few users. For the experiment, we set this limit to 3.

The results of user classification under this unlabeled ads classification scheme is presented in Table 3. Overall, 1,530 and 722 ads were selected from *buy and sell* and *cars & vehicles* categories respectively. We notice that this unlabeled scheme achieves a remarkably close F-measures to those obtained using manually labeled ads data (as reported in Table 2). These results show that a simple semi-supervised setting with only a few n-grams as the initial set can be an effective strategy for user classification without losing much performance.

5. ANALYZING THE POSTING BEHAVIOR

Another hypothesis is that in addition to the text of user ads, the collective posting behavior of the users can convey a lot of information about them. In this section, we exploit this dimension and study the behavior of users using their posting patterns.

As a set of features that describe the posting patterns of the users, we have identified the following:

Posting frequency The frequency with which the users post often provides useful cues as to whether the user is a business. Two frequency-based features that are considered w.r.t time are the average and the standard deviation (SD) of the number of ads per week. The idea behind the former is to separate users by their activity level while the latter

indicates if their posting activity remains consistent over time.

We also consider the inter-arrival time of the ads as an indicator of how actively users utilize the network. The features considered are the average and the standard deviation of the inter-arrival time of the user ads, both in days. The motivation is to identify how soon the users return to list another ad and how consistent are they in such behavior.

As for the distribution of ads in different categories, the standard deviation of the number of ads in different categories is considered. Of course one may count in or out the categories in which a user has no posts. The intuition behind these features is to identify if the user tends to post a large number of ads in a particular category (or set of categories) or if the postings are spread evenly across various categories.

Moreover, it is possible (and quite common) for the users to use the classified ad network consistently for some time (perhaps for some small duration) and then take a long break (possibly weeks or months) before posting ads again. To alleviate the impact of long break times, we divide the active online time of each user into epochs. Within an epoch, the inter-arrival time between any two of his consecutive ads cannot be larger than a week.

Several features over epochs are considered including average and standard deviation of the length of an epoch (in days), the fraction of active time and change in ads per week. The intuition behind the first two features is to identify for how many days the user remains active at a time and how consistent this behavior is. The third feature gives the percentage of overall time during which the user utilizes the network to list an ad. It is calculated as the ratio of the sum of the duration of all epochs (in days) to the total number of days between the date the first ad was posted by the user and the last data collection date. Finally, the fourth feature indicates how the average number of ads per week deviates from the overall trend as compared to the time during which the user is actively utilizing the network. It is calculated for each user as the ratio of the average number of ads in a week over epochs only to the same quantity computed over the entire duration of the dataset. The higher the number is, the more the deviation and vice versa.

Reposts By default, the ads returned for a search request are ordered by the time posted in a decreasing fashion. Thus the newest ads are listed on top and are more visible to the public. A large number of users often repost their ads prematurely to increase their visibility. To capture this posting behavior, we consider for each user the fraction of reposts in the collection of all the ads that the user has posted and the fraction of unique ads that are reposted by the user.

We identify *reposts* using a shingling technique [Broder, 1997], where given an ad, we extract unique 1- and 2-shingles from its titles and description. Then, the resemblance r of two ads A and B can be computed using the Jaccard similarity coefficient as:

$$r(A,B) = \frac{\mid S(A) \cap S(B) \mid}{\mid S(A) \cup S(B) \mid}$$

where $S(A)$ and $S(B)$ are the set of 1- and 2-shingles of ads A and B respectively.

We determine the resemblance of a pair of ads posted by the same user in terms of both its title and description. If either of the resemblance scores for a pair of ads is greater than or equal to 0.8 (determined to be an effective score during our manual checking), we tag the later ad as a repost.

Length features and wanted ads As length based features, we use the average length of ads description both in characters and in words. We also use the fraction of wanted ads to learn more about the usage patterns of the users.

	Business	**Non-Business**
Cluster 1	44 (2.66%)	1,609 (97.34%)
Cluster 2	14 (2.33%)	586 (97.67%)
Cluster 3	26 (4.11%)	607 (95.89%)
Cluster 4	26 (13.06%)	173 (86.94%)

Table 4: Confusion matrix of the clusters and the breakdown of the users in each cluster

5.1 Evaluation and Analysis

Posting patterns are meaningful if the user has more than one post, hence we limit this study to those users who have at least two postings. This reduces the number of qualified annotated users to 3,254 which includes 110 business users (3.38%), 2,975 non-business users (91.43%) and 169 (5.19%) unknown. This dataset is more unbalanced than our ads dataset and we did not expect our classifiers to perform well here. Our experiments confirmed the same; using the RUSEC approach to balance the training data helped to predict more business users but the performance was still worse than those obtained through the ad classification.

To better understand common usage patterns and possibly explain our classification results on behavioral features, we did cluster the users based on the top-10 features. This was done using the Expectation Maximization, which returned four clusters with the confusion matrix shown in Table 4. Studying the distribution of features among clusters revealed that there are significant differences in the posting behavior of users in different clusters. As shown in Table 5, Cluster 1 has the least active users (*inactive* cluster) based on our posting frequency features. The users making up this cluster, on average, remain active for only 1-2 days at a time. In addition, we determined that only 55% of them have posted more than two ads in total, which is in stark contrast to the other groups where at least 90% of the users have done the same. Not surprisingly, Cluster 4, with the largest percentage of business users in its composition, consists of users who use Kijiji most frequently (*active* cluster). Users in the other two groups exhibit values between these two extremes, however, those in Cluster 3 are slightly more active than their counterparts in Cluster 2.

Moreover, we observe that ad description lengths and reposting behavior do not always follow the activity trends mentioned previously. Specifically, *less active* group exhibits negligible reposting activity and have the most succinct ad descriptions on average. Similarly, both *inactive* and *active* users tend to give nearly the same level of details when posting an ad even though the latter has a more prevalent reposting behavior.

Revisiting user classification A manual examination of the users in each cluster reveals some interesting patterns. We noticed that the majority of *inactive* business users do not use Kijiji normally for the promotion of their products and/or services. However, on rare occasions, they have to dispose off some items from their inventory urgently, hence,

	Cluster 1 *(inactive)*	**Cluster 2** *(less active)*	**Cluster 3** *(active)*	**Cluster 4** *(highly active)*
Posting Frequency	Low	Moderate	High	Very High
Reposts	Moderate	Low	High	Very High
Length	Moderate	Low	Moderate	Very High

Table 5: Characteristics of the clusters

are utilizing Kijiji to announce special promotions and discounts. Other types of business users found in this group are also characterized by low activity; examples are small businesses who should be promoting their offerings in *services* category (reserved for small businesses) but list a few of their ads in other categories for additional publicity, enterprises going out of businesses and posting an ad to attract new potential owners, businesses that provide a summary of their services and to promote their official web presence, businesses with time-bound advertisements like a summer camp inviting registrations.

It is not surprising to find that the *highly active* cluster contains the largest percentage of business users in its composition. Some examples of businesses here are those selling cell phone protective cases, providing fresh seed mixes, offering computer repair and wall mounting services, etc. In all these cases, we observed that business users generally do not list a distinct ad for each item or item type they have in their inventory or the different kinds of services offered, but post a general ad detailing their offerings and repost it over time with minor modifications. Accordingly, we found that most of the businesses in this category are *service-oriented*. Likewise, most non-business users in this cluster have only a few items for sale (even one) but they tend to repost their ads often to increase their visibility in the hope of selling their items quickly.

Less active users are characterized by scanty reposting and terse ad descriptions. Thus it is not unexpected that the cluster contains the least fraction of business users in its composition. A manual study of the businesses in this cluster divulged that majority of them can be divided into two categories: (1) individuals providing services and (2) users tagged as business (by annotators) based on the homogeneity in the type of items they listed. The fact that most of these businesses also use Kijiji for their personal needs, i.e., post a sizeable number of ads not related to their business (as reflected by low values of frequency features w.r.t category) and exhibit other non-business like characteristics as mentioned above indicates that many of them are operated on a part-time or seasonal basis.

Finally, *active* users are distinctive since they not only have a considerable number of items to sell, but they also strive to increase the viewership of their ads via reposting. Accordingly, the businesses in this category are established ones like car dealers, heavy equipment sellers, contractors, etc. We observed that unlike *highly active* group, which is dominated by service-oriented businesses who place emphasis on reposting a limited number of ads over time, this cluster is influenced by *product-based* businesses, who generally post separate ads for different kinds of items/services offered. However, a few service-providers are also grouped here because they tend to make minor modifications to the content of their ads when reposting, due to which our heuristic is not able to detect them as reposts. Like *less active* users, a small number of businesses also seem to use Kijiji for their personal purposes, however, the number of such non-business ads are comparatively low.

We conclude our discussion by saying that although certain user classes have more affinity to a particular usage pattern, it is not exclusive to that class only since a significant fraction of members of the other class also manifest the same trend. For the same reason, the collective behavioral features alone are unable to distinguish between the users belonging to the two classes {*business, non-business*} satisfactorily; it remains an open challenge to achieve an adequate separation between the user groups based on the posting patterns of the users. However, analyzing users by different usage patterns, we were able to characterize various kinds of business users which helps us gain a better understanding of how they utilize a classified ad network.

6. CONCLUSIONS

We studied the problem of characterizing users in a classified ad network and detecting if a user is using the network for business. Our experiments on the users of the Kijiji network revealed that the content of the ads posted by a user can provide important clues on detecting business users. We experimented with various content-based features and classification algorithms and analyzed their performance. We further studied the impact of the profile size on the performance and developed a simple semi-supervised strategy to address the cases where labeled data is not available. We also studied the posting behavior of the users and identified four distinct usage patterns that better characterize different classes of business users.

Acknowledgments

This research is supported by the Natural Sciences and Engineering Research Council of Canada.

References

K. Aas and L. Eikvil. Text categorisation: a survey. *Norwegian Computing Center, Technical Report*, 941, 1999.

F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *User Modeling, Adaption and Personalization*, pages 1–12. Springer, 2011.

F. Abel, E. Herder, G.-J. Houben, N. Henze, and D. Krause. Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction*, 23(2-3):169–209, 2013.

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng. (TKDE)*, 17(6):734–749, 2005.

F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *Proc. of the ACM SIGIR Conference*, pages 620–627. ACM, 2009.

F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Proc. of the CEAS Conference*, volume 6, page 12, 2010.

M. W. Berry and M. Castellanos. Survey of text mining. *Computing Reviews*, 45(9):548, 2004.

A. Z. Broder. On the resemblance and containment of documents. In *Proc. of the Compression and Complexity Conference*, pages 21–29, 1997.

F. Carmagnola, F. Cena, O. Cortassa, C. Gena, and I. Torre. Towards a tag-based user model: How can user model benefit from tags? In *Proc. of the User Modeling Conference*, pages 445–449. Springer, 2007.

C. Chen, K. Wu, V. Srinivasan, and X. Zhang. Battling the internet water army: Detection of hidden paid posters. In *Proc. of the ASONAM Conference*, pages 116–120, 2013.

A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine learning*, 2(2):139–172, 1987.

W. Klement, S. Wilk, W. Michaowski, and S. Matwin. Dealing with severely imbalanced data. *Proc. of the PAKDD Conference*, page 14, 2009.

J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *Proc. of the IUI Conference*, pages 31–40, 2010.

A. Makazhanov, D. Rafiei, and M. Waqar. Predicting political preference of twitter users. *SNAM*, 4(1):1–15, 2014.

A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems (NIPS)*, 2:849–856, 2002.

J. Orwant. Heterogeneous learning in the doppelgänger user modeling system. *User Modeling and User-Adapted Interaction*, 4(2):107–130, 1994.

G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

K. Schöfegger, C. Körner, P. Singer, and M. Granitzer. Learning user characteristics from social tagging behavior. In *Proc. of the conference on Hypertext and social media*, pages 207–212. ACM, 2012.

M. Shmueli-Scheuer, H. Roitman, D. Carmel, Y. Mass, and D. Konopnicki. Extracting user profiles from large scale data. In *Proc. of the MDAC Workshop*, page 4, 2010.

J. Stoyanovich, S. Amer-Yahia, C. Marlow, and C. Yu. Leveraging tagging to model user interests in del. icio. us. In *AAAI Symp.: Social Information Processing*, pages 104–109, 2008.

A. Sun, M. A. Suryanto, and Y. Liu. Blog classification using tags: An empirical study. In *Pro. of ADL Conference*, pages 307–316. Springer, 2007.