Characterizing Users and Tracking their Activities in Online Classified Ads

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

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

Characterizing users and tracking their activities in online classified ads is a topic of great importance. However, some of the underlying problems associated with modeling users and detecting their behavioral changes have not been well-studied.

In this paper, we develop a probabilistic framework for characterizing users and quantifying some of the spatial and temporal variations in their posts. Our work on characterizing users study the problem in the context of detecting if a user belongs to a class, based on the ads the user has posted. Our approach is based on user profiling, where given statistics on user posts, the affinity of a user to a class is estimated. We show how profiles can be constructed with and without training data and report the effectiveness of our approaches in detecting two user classes *business* and *non-business*.

Our work on quantifying changes due to spatial and temporal variations is based on a probabilistic model of user behavior and a generative model that can predict ad posts from each location. We evaluate these models on a relatively large set of users and ads, and report our results on two classes of users monitored over a period of almost a year.

Keywords: classified ads; user modeling; user profiling; change detection; behavioral analysis.

1. Introduction

User 64073399 has posted a classified sales ad for a 2008 Honda Accord, and the ad claims to be "for sale by owner." Can this claim be validated? Is there any other information source to verify this claim? A user moves to a new location; are we expecting changes in his posting behavior? Or in general, how can we quantify temporal and spatial changes in user postings? Are some ad types expected in certain locations or times? These are all important questions but difficult to answer for print ads due to the sparsity of the data and the lack of meta data. However, with many individuals and businesses are now moving away from traditional print ads and more towards online classified advertising, there are more opportunities for user modeling and analysis; this paper explores some of those opportunities.

Classified ad sites have some commonalities with e-commerce retail sites such as Amazon and AliExpress and marketplaces such as Ebay, but they are different. The former provides a way to list items, services, community events or properties for sale with a focus on selling locally in the community. On the other hand, the latter, focused on connecting consumers from all over the world to sellers, typically places a greater emphasis on the satisfaction of their users by increasing the quality and the reliability of their services as well as protecting users from scams. Therefore, they often charge a fee for listing an ad, allow buyers to view a detailed transaction history of the seller and incorporate a feedback system whereby the buyers rate a seller after the completion of a transaction.

In the absence of rigorous checks and controls on classified ad users, ascertaining if a user of a classified ad site belongs to a class involves various challenges. For example, consider the two classes *business* and *non-business*.

Although classified ads are mostly placed by private individuals, many businesses also use the medium for the promotion of their products and/or services, due to the low cost and the large customer base. However, 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 as noted above, 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 desirable consequences. For example, in a system that traditionally involves no user feedback, it gives the buyers better information about the nature of the seller. The government may use the data to better regulate certain sectors of economy which may not be reachable otherwise. An online ad site 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.

Problem Statement Can the class of a user be detected from the posted ads, and how user classes and characteristics change with temporal and spatial variations?

Our hypothesis is that certain traits of a user can be learned from the language model of the ads the user posts and that temporal and spatial variations influence some of the traits.

A more concrete problem is detecting if a user belongs to one of the two given classes; *business* (using the network for promotion of his enterprise) or *non-business* (using the network for personal use). 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. 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. Our approach for distinguishing

between the two user classes as mentioned above mostly relies on building language models of both classes and determining if an ad belongs to a particular class based on the mentions of terms from the language models.

A related problem studied in this paper is tracking changes in user traits and possibly predicting them; such predictions can be an important information for applications that are built on top and may also help in better understanding of the marketplace dynamics in classified ads. For example, less changes are expected in the profile of more established or specialized users and this may help identify such users. Similarly changes in seasonal patterns and ad distributions may reveal demographic changes of different neighborhoods. There is a growing body of work on tracking topics,[?] quotes, and news pieces[?,?] as they evolve or spread over time. There is also past work on user modeling in general and activity tracking in particular, such as navigation between pages.[?,?] However, classified ads are somewhat different from both information networks (e.g. the Web) and social networks (e.g. Facebook, Twitter, etc.) in that there is no direct tie between social actors (e.g. users or ads), and the social acts (if they can be called so) are in the form of weak ties (e.g. users who post to the same category). We are not aware of much study of the aforementioned issues in the context of classified ads; more specifically we could only find a limited number of work studying the social and economic impact of online classified ads (e.g. the work of Kroft et al.[?]).

To track such changes, we present a probabilistic model of user behavior based on the interactions between users, ads and post categories. The user behavior can change over time and those changes are often triggered by external forces; we show how changes can be tracked and the collective behavior of users with similar interests can be identified. We further study the changes in behavior due to spatial variations and show that how a generative model may predict ad production based on the location from which the ad is posted. We evaluate our models through experiments conducted on data collected from a real classified ad site, and report some of our results and findings.

Our contributions can be summarized as follows: (1) we study the problem of profiling classified ad users, tracking their activities and the marketplace dynamics; (2) we present a method for detecting business users based on their ads; (3) we show how some of the distinctive usage patterns of a desired user class can be detected; (4) we present a probabilistic model of user behavior which allows us to analyze the relationships between users, ads and categories and detect temporal and spatial variations; (5) we evaluate our models and algorithms using a crawl of the ads posted to a classified ad site over a year and report some of our findings and analysis as well as comparisons to a few competitive baselines.

Organization The rest of this paper is organized as follows. We review the related work in Section 2. Section 3 details our methodology for classifying users using ads content and corresponding results in a supervised as well as semi-supervised setting. The posting behavior of the users is analyzed in the same context in Section 4. In Section 5, we both model and analyze the network from various dimensions

using the results of user characterization obtained previously. The paper is summarized and concluded in Section 6, where we also present various avenues for future research.

2. Related Work

Our work in this paper relates to three lines of research: (1) *text and short text classification*, (2) *tracking user posting behavior*, and (3) *modeling spatial and temporal variations*. The literature on online classified ads mostly cover social and economical impacts, such as the impact of online ads on local newspapers[?,?] and linking the ads to sexually transmitted diseases.[?,?,?] However, we are not aware of any studies on characterizing users of a classified ad site or a quantitative analysis of their behavior on a large scale.

2.1. *Text and short text classification*

Our work utilizes the content of an ad to determine its affinity to *business* and *non-business* classes, hence it is related to the body of work on text and short text classification. Several techniques have been developed for text classification, and a relatively comprehensive survey of them can be found in the works of Aas and Eikvil[?] and Berry and Castellanos.[?] Examples of topical text classification can be found in spam email detection,[?] classifying news stories[?] and blog posts,[?] etc.

The related work on short text classification is maybe more relevant; this line of work often uses, in addition to text, meta information from other sources such as Wikipedia and WordNet[?,?] to classify tweets. Sriram et al.[?] use author-specific features, instead of meta information from other sources, to classify text to some genetic classes including news, opinion, deal, private; their work shows that high classification accuracies can be achieved. More recently topic models with word embeddings are as well applied to short text classification to improve the accuracy.[?] For classified ads, however, there is not much meta information about users from other sources, and it is not clear if the semantics of the words in news will transfer well to classified ads (though this area is an interesting subject for further studies). A closely related work to ours is that of Makazhanov et al.[?] 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 users directly; instead, we classify the ads, which form the basis for classifying users.

2.2. *Tracking user posting behavior*

Our work builds models of users based on not only the content of their ads but also their posting patterns such as time of the posts, posted location (if available), etc., hence it is related to similar modeling exercises in social media. Related work includes the literature on characterizing the behaviour of online social network users to learn global activity patterns such as daily and weekly patterns and content

distribution workload.[?,?] Xu et al.[?] identify three factors influencing a user post: (1) breaking news, (2) friends' posts, and (3) user interest; accordingly, they propose a mixture topic model to analyze user posting behavior. Li et al.[?] study the profiling of users' home location based on *following* and *tweeting* relationships between users.

Content-based models of users are more relevant to our work since not much is known about a classified ad poster other the posted ads. Liu et al.[?] 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. Nichelson and Macskassy[?] and later Shen et al.[?] learn users' topics of interest on Twitter based on the entities they mention in their posts. Weng et al.[?] identify the topics of interest of a Twitter user by collapsing the tweets of a user into a big document and learning a Latent Dirichlet Allocation (LDA) model. Some of these studies are related to the extensive body of work on recommender systems, which learn a model of users' interest based on their past behavior. A comprehensive survey of such works, covering the work before 2005, has been put forward by Adomavicius and Tuzhilin.[?]

Extensive work has been done on leveraging user models to find other users with similar behavior. Benevenuto et al.[?] 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,[?] commercial campaigns in Community Question and Answer (CQA) forums[?] and spammers as well as content promoters in online video social networks.[?]

2.3. *Modeling spatial and temporal variations*

The spatio-temporal dynamics of terms and topics is studied in the context of search engine queries,[?] Twitter hashtags,[?] etc. Beitzel et al.[?] examine a query log with billions of web queries issued over a period of six-months, categorize them into topics and analyze the trends in category popularity over time. Their findings suggest that some categories change more drastically than others over both short-term periods (e.g., hours, days) and in long-term (e.g., months, seasons). Our work is related to some of these studies in that we also analyze the classified ad network from various dimensions (category, time, location) in order to gain a better understanding of ad posting dynamics. Our modeling approaches are similar; for example we model the ad posts in categories as Bernoulli trials, similar to the approach taken by Backstrom et al.[?] in the context of Web queries.

3. From Ad Content to User Classification

We study the problem of user classification by considering the content of the ads posted by users. To motivate this approach, let us consider the two ads shown in Figure 1, and suppose that these are the only ads posted by their respective posters. Going through the ad descriptions and with no other information about these users,

one may easily tag the user who posted the ad in Figure 1a as a business user whereas the poster of the ad in Figure 1b is very likely to be labeled as non-business. By analyzing the decisions, it becomes apparent that the text of user postings provides important clues regarding the classification of the user. For example, we do not anticipate common usage of expressions such as *in our family's handmade business*, *we ensure*, *we create*, *amazing prices*, *starting from* by non-business users. Likewise, we also do not expect many business users to use the phrases *just upgraded*, *make me an offer* etc. in their postings.

In light of this motivation, the problem of user classification is linked to the classification of individual ads of the user. In this section, we study the latter problem, i.e.

> Given an ad a, predict if a has more affinity for *business* class or *non-business* class.

Later in the section, we will use the results of ads classification to come up with a label for the user.

We represent each class label as an abstract concept and with it we associate a ranked list of weighted terms. We call this the *profile* of the class. An ad may *mention* terms from the profiles of both classes. Using various statistics of these *mentions*, we seek to characterize an ad as being business oriented or non-business oriented.

3.1. *Building Profiles*

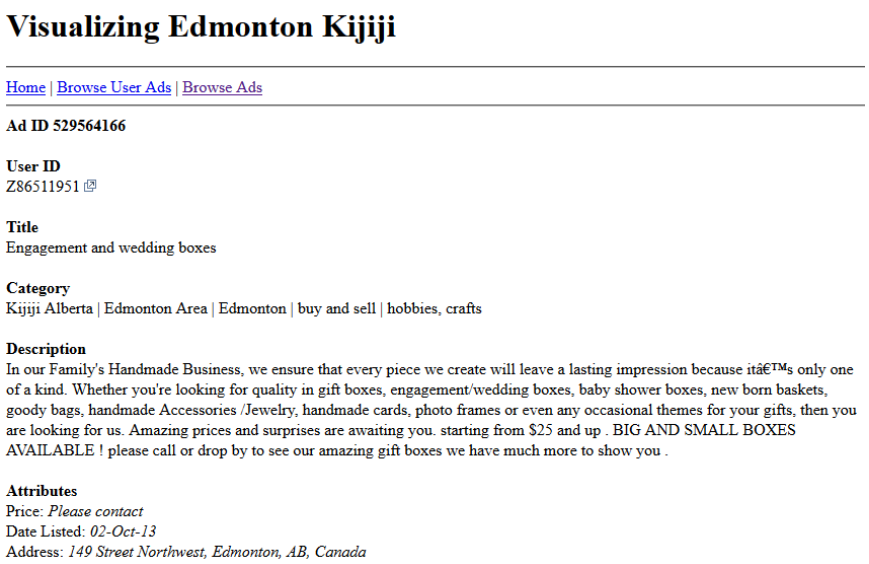
Let l denote the class label of an ad that we want to predict. In our case, l may be one of the elements in the set $L = \{business, non\text{-}business\}$. Let D_l denote the collection of all ads with class label l. We refer to this as the class corpus. The entire corpus, therefore, is denoted by $D = \{D_l \mid \forall l \in L\}$ and its vocabulary is represented as V.

We build a language model (LM) for each class as well as one for the entire corpus, and calculate the Kullback-Liebler (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.[?]). The divergence score of an individual term can then be used as a measure of the importance of the term to a specific class. This provides a ranked and weighted list of class-specific terms.

tf-idf scores are used to calculate term probabilities for a particular corpus. For the entire corpus, 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



Visualizing Edmonton Kijiji

Home | Browse User Ads | Browse Ads

Ad ID 529564166

User ID
Z86511951

Title
Engagement and wedding boxes

Category
Kijiji Alberta | Edmonton Area | Edmonton | buy and sell | hobbies, crafts

Description
In our Family's Handmade Business, we ensure that every piece we create will leave a lasting impression because itâ€™s only one of a kind. Whether you're looking for quality in gift boxes, engagement/wedding boxes, baby shower boxes, new born baskets, goody bags, handmade Accessories /Jewelry, handmade cards, photo frames or even any occasional themes for your gifts, then you are looking for us. Amazing prices and surprises are awaiting you. starting from $25 and up . BIG AND SMALL BOXES AVAILABLE ! please call or drop by to see our amazing gift boxes we have much more to show you .

Attributes
Price: *Please contact*
Date Listed: *02-Oct-13*
Address: *149 Street Northwest, Edmonton, AB, Canada*

(a) Business Ad

Visualizing Edmonton Kijiji

Home | Browse User Ads | Browse Ads

Ad ID 465658229

User ID
Z73506402

Title
27" tv

Category
Kijiji Alberta | Edmonton Area | Edmonton | buy and sell | electronics

Description
Not a flat screen. Excellent condition, hardly ever used. But it do work perfectly. Just upgraded and want to get this from under my feet. make me an offer

Attributes
Date Listed: *16-Mar-13*
Price: *Please contact*
Last Edited: *17-Mar-13*
Address: *West Edmonton, Edmonton, AB, Canada*

(b) Non-Business Ad

Fig. 1. Sample Business and Non-Business ads.

and $udf(t, D) = df(t, D)/|D|$ represents the probability of t appearing in a document in 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.

It may be the case that certain terms present in V are not represented in the class corpus. To account for these missing terms, their 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:

$$I(t, l) = P(t \mid D_l) \ln \frac{P(t \mid D_l)}{P(t \mid D)}. \quad (1)$$

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. With a ranked and weighted list of terms for each class, each class profile can be built by selecting top-N terms from this list.

Table 1 lists some of the top-ranked bigrams from the class profiles of categories in the abridged dataset (Section 3.3). We observe that businesses (across both categories) tend to focus on first-person plural pronouns (we, us, our), a trend also shown by Packard et al.[?] On the other hand, non-business users are more likely to mention first-person singular pronouns (I, me, mine) in their ads. Note that the first few bigrams in non-business profiles of both the categories are the same. This is due to the fact that they are extracted from a sentence that is automatically appended at the end of the ad description if the user is posting the ad through one of Kijiji's mobile applications. This indicates that non-businesses use smartphone applications to post classified ads on Kijiji much more extensively than the business users.

Table 1. Top ranked bigrams from the class profiles.

(buy and sell)		*(cars & vehicles)*	
Business	**Non-Business**	**Business**	**Non-Business**
we have	posted with	information on	was posted
for more	kijiji mobile	see more	posted with
http www	mobile app	of our	kijiji mobile
selection of	was posted	on kijiji	mobile app
for each	i have	contact information	i have
we are	or text	we are	brand new
hours monday	excellent condition	our dealership	i am
visit our	i am	call email	comes with
call us	comes with	serve you	selling my
please visit	pick up	our website	like new

3.2. *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 (assuming that an ad cannot belong to both *business* and *non-business* classes). 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 *unknown* and are ignored for that particular class in the ads classification task. In the presence of multiple classes, one can simply construct a binary classifier for each class.

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.[?]

Class profiles were built for each top-level category in which an ad is posted; we expected the same or similar class terms for the 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*). For example, *financing* may be a popular term with the car dealers but is not often used by businesses in other categories. Likewise, *re/max* and *registered breeder* are expected to pop up frequently in business ads of *real estate* and *pets* categories respectively (not considered in this work) but might not be very popular with other businesses. However, within a top-level category, the general vocabulary of business and non-business users should not differ very much owing to the categorization in online classified ad websites.

3.3. *Experimental Setup*

Our experiments were conducted on datasets from two independently-run classified ad sites Kijiji[a] and Craigslist[b]. Kijiji is a popular online classified ad service that allows users to post free classified ads in different categories. The ad service was chosen because of its popularity in Canada, its wide presence in multiple other countries and the fact that it allows searches for other ads of a user. The latter point allows grouping the postings of a user and build models for tracing their activities. Craigslist was chosen because it is reported to be the largest classified ad site with presence in 70 countries. Unlike Kijiji, Craigslist does not necessarily allow searches for other ads of the poster.

Collecting Kijiji dataset We built a crawler to extract the ads from Edmonton Kijiji [c] which services the cities of Edmonton and St. Albert as well as the nearby Strathcona County. The crawler was run once every day over a period of nine months from May 2013 to January 2014. During the crawling period, all posted ads that were active at the time of the crawl were fetched (ignoring the ones that have previously scraped) and various fields including ad ID, title, category, listing date, address (if present), price (if present), description (in plain text) and user id were extracted.

The dataset had some very diverse categories, such as *buy and sell* which allows users to purchase or offer for sale items ranging from books to entire businesses. On the other hand, many categories were relatively limited in their scope, such as *pets*. For nearly all the experiments in this paper, we utilize data from one representative of both groups. Specifically, we selected *buy and sell* category from the diverse bracket, since a manual examination revealed that it presents the highest nature of imbalance in terms of *business* and *non-business* classes, thereby, making the user classification task most challenging. Moreover, we chose *cars & vehicles* from more specialized categories. These are the two largest categories in our dataset in terms of the number of ads, as shown in Figure 2. We refer to these two categories as our *abridged dataset*. Table 2 reports some statistics for the complete as well as the abridged dataset.

Table 2. Statistics for complete and abridged datasets.

Statistic	Complete dataset	Abridged dataset
Number of ads	3,420,050	2,540,316
Number of users	410,637	318,672
Minimum ads per user	1	1
Maximum ads per user	4,842	4,842
Average ads per user	8.33	7.97
Median ads per user	2	2

[a] `http://www.kijiji.ca`
[b] `http://www.craigslist.ca`
[c] `http://edmonton.kijiji.ca`

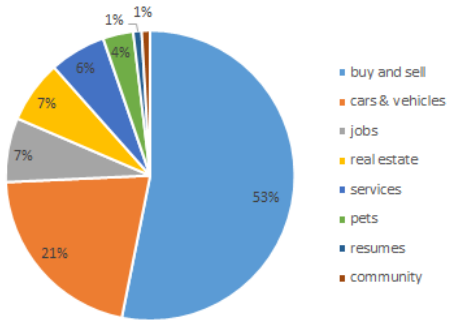


Fig. 2. Distribution of the ads in various categories for the Kijiji dataset

Collecting Craigslist dataset To validate our basic approach, we also collected ads from Edmonton Craigslist[d] during December 2016. Unlike Kijiji, classified ads related to automobiles are not listed in a separate top-level category for Craigslist. Instead the various vehicle categories are merged into a general *for sale* category. Thus, the dataset was post-processed to separate ads into *buy and sell* and *cars & vehicles* categories to be homogenous with the one from Kijiji network. In total, a little over a 100 ads for each of the aforementioned categories were collected.

Ground truth To prepare the ground truth for our experiments, a random sample of the entities in the dataset was manually labeled into *business* and *non-business* classes. Since distinguishing between the two classes is sometimes confusing, annotators had the additional option to mark an entity as *unknown*; these were ignored in our experiments.

Table 3. Statistics for Kijiji dataset prepared for ad classification task.

	Total	**Business**	**Non-Business**	**Unknown**
(buy and sell)				
Number of ads	1,858	150	1,585	123
Percentage of ads	-	8.07	85.31	6.62
(cars & vehicles)				
Number of ads	756	150	578	28
Percentage of ads	-	19.84	76.45	3.7

Table 4. Statistics for Kijiji dataset prepared for user classification task.

	Total	**Business**	**Non-Business**	**Unknown**
Number of users	5,000	157	4,634	209
Percentage of users	-	3.14	92.68	4.18

Tables 3 and 4 show some statistics of the labeled data from Kijiji network

[d]`http://edmonton.craigslist.ca`

for the classification of ads and users respectively. From the Tables, it is quite clear that while both labeled datasets suffer from class imbalance, the degree of the imbalance decreases as we go from users to ads. This is not very unexpected, since we would expect most business users to post ads regularly promoting their offerings as compared to non-business users who would post as needed. Moreover, the percentage of business ads in the *cars & vehicles* category is more than twice of that for the *buy and sell* category. This indicates that automobile businesses are more proactive (either a larger number of them use Kijiji to promote their business or they post more ads per user or both) than their counterparts in the *buy and sell* category.

Similar statistics for equivalent tagging of ads collected from Craigslist are reported in Table 5.

Table 5. Statistics for Craigslist dataset prepared for ad classification task.

	Total	**Business**	**Non-Business**	**Unknown**
(buy and sell)				
Number of ads	111	13	87	11
Percentage of ads	-	11.71	78.37	9.92
(cars & vehicles)				
Number of ads	105	32	68	5
Percentage of ads	-	30.48	64.76	4.76

3.4. *Ad Classification results*

In a 10-fold cross-validation experiment, we evaluated the performance of our ad classification methodology, in terms of Precision (P), Recall (R) and F-measure (F) scores. We experimented with three classifiers: decision tree based Random Forest (RF), support vector machine backed Sequential Minimal Optimization (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 under-sampling with an ensemble of classifiers (RUSEC) and (2) using the imbalanced data (IMB) as is. For random under-sampling, each classifier in the ensemble of three is trained on a balanced sample of the 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 as done in the literature.[?] 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 ignored the numbers and rare terms (those occurring in two or less ads).

Table 6 shows the results of ads classification on the Kijiji dataset. It can be noted that 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

Table 6. Results of the ad classification for Kijiji dataset with the profile size set to 100.

		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

trend is to be expected. Since the training data in the IMB approach is imbalanced, the respective classifiers tend to maximize their overall accuracy, and this leads to them optimizing predictions for the dominant class. Due to this behavior, the minority class (*business*) suffers in recall. 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 are trained on the balanced data, thus, they 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* in comparison to *cars & vehicles*. The trend is much more noticeable when considering the RUSEC classifiers. We believe this is because of the fact that *buy and sell* is a much more diverse category. Hence, 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*. Again, this trend is more apparent for RUSEC classifiers which achieve nearly similar recall for both categories. 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* (Table 3). Therefore, a similar percentage of non-business instances being misclassified has a stronger impact on the precision of business users for *buy and sell* than *cars & vehicles*.

To evaluate the generality of our models and algorithms and to investigate any possible bias towards a specific online classified ad service, we evaluated our approach in the context of another popular classified ad portal, Craigslist. For the experiment, we treated the labeled ads dataset from Kijiji (Table 3) as a training set and used the classifiers trained on this set to classify ads gathered from Craigslist (Table 5). The results, as shown in Table 7, reveal the same trends as discussed previously for the Kijiji dataset. Interestingly, Logistic Regression seems to

perform best for RUSEC training approach as well. The precision for business class for *buy and sell* category has markedly increased due to less data imbalance as well as increased recall values for the non-business class. Accordingly, we notice comparable or even better F-measures for Craigslist in both categories. This confirms the generality of our method and that models learned from ads in one ad service nicely transfer to other ad services. This was expected since (1) our models do not use any site-specific tags or features, and (2) the main characteristics of a classified ad (such as writing style and the language model) is not expected to change much between different ad services and even between online and print ads.

Table 7. Results of the ad classification for Craigslist dataset with the profile size set to 100.

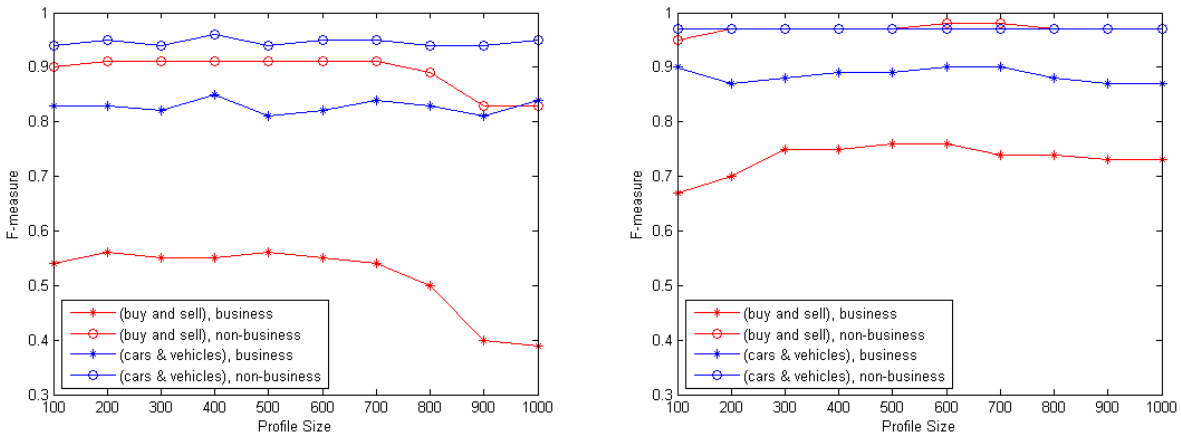
		Business			**Non-Business**			**Business**			**Non-Business**		
		P	R	F	P	R	F	P	R	F	P	R	F
		buy and sell						*cars & vehicles*					
RUSEC	LR	0.71	0.92	**0.8**	0.99	0.87	**0.93**	0.85	0.91	**0.88**	0.95	0.9	0.92
	SMO	0.63	0.92	0.75	0.99	0.86	0.92	0.78	0.91	0.84	0.95	0.88	0.92
	RF	0.6	0.92	0.73	0.99	0.85	0.91	0.91	0.94	0.92	0.97	0.96	**0.96**
IMB	LR	1.0	0.69	**0.82**	0.95	0.97	**0.96**	1.0	0.88	**0.93**	0.94	0.99	**0.96**
	SMO	1.0	0.69	0.82	0.95	0.97	0.96	0.93	0.81	0.87	0.92	0.97	0.94
	RF	1.0	0.62	0.76	0.95	0.97	0.96	0.91	0.66	0.76	0.87	0.97	0.94

3.4.1. *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) for Kijiji dataset and repeated the same experiment by increasing the profile size from 100 to 1,000 with an increment size set to 100.

Figure 3 shows the F-measures for both business and non-business classes in each category and for each method. 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. However, as we continue to increase the profile size, scores become stable (meaning that the newly added terms have a negligible impact on results) and even start decreasing. This trend is not surprising, since the lower the terms are in the profile, the more they are a part of common users vocabulary rather than being distinctive for a particular class (Section 3.1).

The only exception to the aforementioned trend is the IMB method on the *(buy and sell)* category. Here, as we increase the profile size, the F-measure increases significantly (nearly 8%) for the first two iterations, after which the scores become stable and even show a slight decline as more terms are incorporated into the profile.



(a) RUSEC (b) IMB

Fig. 3. Ad classification performance varying the profile size.

3.4.2. *Feature analysis*

To rank the features, we set the profile size at 200 and ordered the features based on their information gains (as computed in Weka). We performed a cross-validated ranking, in which the ranks of all features were averaged over the number of folds (10 in our case). Table 8 lists the top-10 features for both categories. Note that the features tagged with R and D are relative values whereas those under T are absolute values. For example, consider the *number of mentions* of a term t. The R feature gives the ratio of the number of ads in class that mention t and the number of ads in the population that mention t. The D feature is the difference between the number of mentions of t in class and the number of mentions of t in the population.

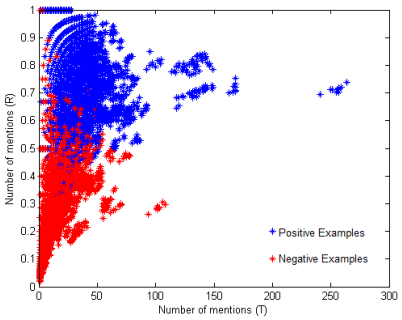
Table 8. Feature ranking for ads classification task.

(buy and sell)			*(cars & vehicles)*		
Feature	**Domain**	**Average rank**	**Feature**	**Domain**	**Average rank**
Number of mentions	R	1 ± 0	Number of mentions	R	1 ± 0
Average weight of mentions	R	2 ± 0	Minimum rank of mentions	R	2 ± 0
Average weight of mentions	D	3 ± 0	Minimum rank of mentions	D	3 ± 0
Average weight of mentions	T	4 ± 0	Number of mentions	D	4 ± 0
Maximum rank of mentions	R	5 ± 0	Minimum rank of mentions	T	5 ± 0
Number of mentions	D	6.1 ± 0.3	Number of mentions	T	6 ± 0
Minimum rank of mentions	R	6.9 ± 0.3	Average rank of mentions	R	7 ± 0
Minimum rank of mentions	D	8 ± 0	Average rank of mentions	D	8 ± 0
Maximum rank of mentions	D	9 ± 0	Maximum rank of mentions	R	9 ± 0
Number of mentions	T	10 ± 0	Average rank of mentions	T	10 ± 0

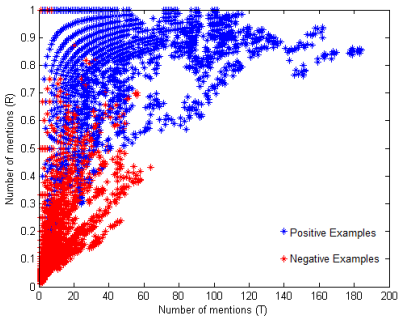
We notice that computing features over domains other than the *T-domain*

turned out to be very effective as at least 7 of the top-10 features are from *R-* and *D-domains* for both categories.

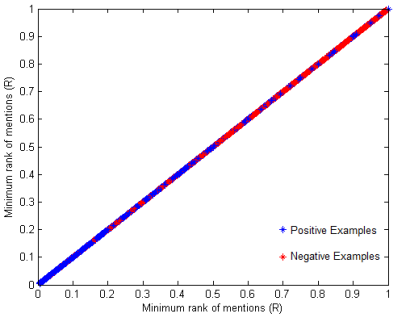
It can be observed from Table 8 that *relative number of mentions* is the most effective feature for both categories, more accurately dividing positive and negative examples. As expected, ads that mention terms from the profile of a particular class more often typically have a greater affinity towards that class. This trend can be viewed in Figures 4a and 4b. 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* is low. However, using *relative number of mentions*, the instances can be differentiated very easily.



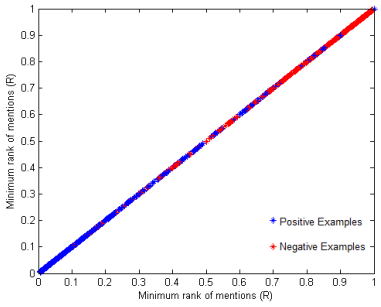
(a) Number of mentions (T) / Number of mentions (R) for *(buy and sell)*



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



(c) Minimum rank of mentions (R) for *(buy and sell)*



(d) Minimum rank of mentions (R) for *(cars & vehicles)*

Fig. 4. Distribution of training examples across different feature spaces.

Likewise, ads that mention top-ranked terms from the profile of a particular class as compared to the other one usually indicate a stronger relevance towards the respective class (Figures 4c and 4d). Accordingly, we find that *relative minimum*

rank of mentions is an important feature for both categories.

3.5. *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 ads of the users in our dataset. We used the 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.
- **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.
- **Weighted Random Baseline (WR)** Finally, we compare our method to a weighted random baseline. For each user, we generate a random real number between 0 and 1. If the number is less than or equal to 0.0314 (the

underlying distribution of business users as shown in Table 4), we classify the user as business. Otherwise, the user is classified as non-business.

Table 9. User classification performance results with the profile size set to 200.

		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
	WR	0.03	0.03	0.03	0.97	0.97	0.97

The result of the user classification task is presented in Table 9. As discussed in Section 3.4, 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.

Of particular importance is the fact that the precision of *Dealer* baseline is not 1.0 as one would expect. We re-checked the users who self-identify themselves as businesses but were classified as non-businesses by human annotators. All annotators were unanimous on the classification of such users and we found no evidence that a mistake has been made on their part. Specifically, it appeared that all such users had tagged themselves falsely in order to promote their ads and to sell their items urgently.

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 there is very strong evidence against the null hypothesis in favor of the alternative, thereby, the difference in performance is statistically significant.

3.6. *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 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. This limit on the number of ads per user was implicit in our manually labeled ads dataset; since a random sample of all ads was taken, the probability that more than three ads per a user could make into the sample was low (owing to the large number of ads in the collection and the limited size of the training data).

Table 10. User classification performance results using unlabeled data and the profile size set to 200.

		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**

The results of user classification under this unlabeled ads classification scheme is presented in Table 10. Overall, 1,530 and 722 ads were selected from *buy and sell* and *cars & vehicles* categories respectively. We notice that this unlabeled scheme achieves remarkably close F-measures to those obtained using manually labeled ads data (as reported in Table 9).

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.

4. Analyzing the Posting Behavior

We have so far looked at the problem of classifying users into one of the {*business, non-business*} classes based on the contents of their posted ads. Another hypothesis

is that in addition to the text of user ads, the collective behavior of the users in posting ads can reveal more information about them. In this section, we explore this dimension and study the behavior of users using their posting patterns.

4.1. *Behavioral Features*

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

4.1.1. *Posting frequency*

The frequency with which a user posts on a classified ad network 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 or experiences great fluctuations.

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 after posting one already 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 model how closely the ads of a user follow a collective theme by identifying 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.

4.1.2. *Reposts*

By default, the ads returned for a search request on most (if not all) classified ad sites are ordered by the time the ads are posted in a descending fashion. Thus the newest ads 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,[?] 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, and $|X|$ is the size of set X.

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.

4.1.3. *Length features and wanted ads*

As length based features, we use the average length of ads description both in characters and in words. The goal here is to capture the level of details that is used to describe an item or a service. We also use the fraction of wanted ads (i.e., ads in which an item or a service is required, instead of offered) to learn more about the usage patterns of the users.

4.2. *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 the users dataset used for ads classification 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.

Table 11. Confusion matrix of the clusters detected by the EM algorithm and the break-down of the users in each cluster.

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

To better understand common usage patterns and possibly explain our classification results on behavioral features, we clustered the users based on the top-10 features, selected based on their information gain. This was done using the Expectation Maximization, which returned four clusters with the confusion matrix shown in Table 11. 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 12, Cluster 1 has the least active users 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 and remain active for nearly 25% of the time on average. 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.

Table 12. Characteristics of the clusters

	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

Moreover, we observe that ad description lengths and reposting behavior do not always follow the activity trends mentioned previously. Specifically, Cluster 2 exhibits negligible reposting activity and have the most succinct ad descriptions on average. Similarly, users in both Cluster 1 and Cluster 4 tend to give nearly the same level of details when posting an ad even though the latter has a more prevalent reposting behavior.

4.3. *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 (in Cluster 1) 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, 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 etc.

It is not surprising to find that the *highly active* group (in Cluster 4) 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 (in Cluster 2) are characterized by scanty reposting and terse ad descriptions. Thus it is not unexpected that the group 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 (in Cluster 3) 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.

5. Tracking User Activities and Network Dynamics

In this section, we present a probabilistic model of user behavior based on the interactions between users, ads and post categories. We show how changes in the user behavior can be tracked and the collective behavior of users with similar interests can be identified. Moreover, we study the changes in behavior due to spatial variations and posit a generative model for postings based on user locations.

5.1. *Temporal Changes in User Behavior*

A user of a classified ad site can have multiple ads each listed under different categories. Hence the set of categories in which a user posts defines a probability distribution.

Definition 1. (User Profile): Given a set of categories C and a set of users U, let $p_{u,c}$ be the probability that user $u \in U$ posts an ad in category $c \in C$. The posting profile of u can be defined in terms of the distribution of his ads in different categories, i.e.

$$P(u) = \{(c, p_{u,c}) \mid c \in C\}.$$

Since a user profile describes the posting behavior of a user in terms of the categories in which the user is likely to post, the changes in behavior can be analyzed based on the changes to the profile. To quantify temporal changes in user behavior, we place a sliding window over user postings and construct a profile for each window. Treating a profile as a population distribution, we measure the evolutionary distance between two profiles in terms of the change that is necessary to transform one distribution into another. We measure this change in terms of the mean absolute difference of the two distributions; this corresponds to the Manhattan (or L_1) distance which is also used in similar contexts (e.g. in the population changes of alleles[?] and more recently in temporal evolution of Twitter user profiles[?]). It is computed between two distributions $\vec{x}$ and $\vec{y}$ as

$$d(\vec{x}, \vec{y}) = \sum_i |x_i - y_i|.$$

Each profile vector is a probability distribution, hence this distance ranges in [0..2].

In an experiment to validate the proposed model, we studied the changes in the profiles of business and non-business users. The experiment included every user

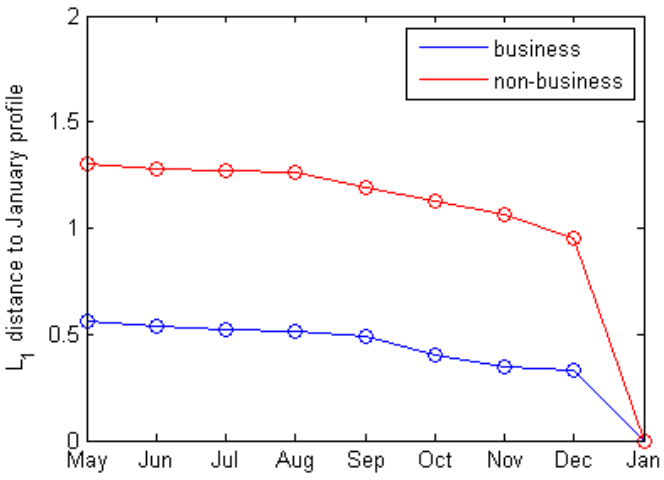


Fig. 5. Temporal changes in user profiles.

who posted at least one ad every month over the duration of our dataset (May 2013 - January 2014). For each user, we assessed the change in profile in terms of the Manhattan distance between the profile for the corresponding month to that of the last month of our dataset (January 2014). The results are shown in Figure 5. As expected, the distance is zero for the last month since the profiles that are being compared are the same. We notice that non-business users exhibit continuously stronger changes in their profiles than business users, as revealed by changes in the L_1 distance. This result is expected, and it conforms to our intuition that private individuals tend to use the classified ad network only when a particular need arises, and as a result, their ads are likely to be scattered in different categories based on the nature of their needs at the time. On the other hand, businesses use the network for advertising their products (or services). Hence, we would expect them to be consistent in their postings in the categories related to their enterprise.

We also observe that for both user groups, the L_1 distance gradually decreases over time; this is simply because the user needs (and consequently their posting behaviors) are less likely to change in shorter time intervals.

5.2. *Distinctive Categories for User Groups*

Given a set of users with possibly some commonalities (e.g. a special interest group), we want to know some of the distinctive features in terms of the posting behavior shared by the members. Specifically, in this case, we examine the categories that are the most distinctive for a particular user group.

For this purpose, we use the same methodology that we adopted when building the profiles of our user classes in Section 3.1. We use the KL-divergence and compute

the importance score of an ad category c for a particular class label l as

$$I(c,l) = P(c \mid U_l) \ln \frac{P(c \mid U_l)}{P(c \mid U)} \tag{2}$$

where U represents the entire corpus i.e., the collection of all users and U_l denotes users belonging to class label l.

The marginal probabilities of categories for the entire corpus $P(c \mid U)$ and class corpora $P(c \mid U_l)$ are calculated using a *tf-idf* weighting scheme similar to the one detailed in Section 3.1; the difference is that $\overline{tf}(c,U)$ now represents the average number of ads posted in category c by the users in U and $df(c,U)$ denotes the number of users in U who posted an ad in category c. The categories having larger values for Equation 2 are the most distinctive for the respective user group (U_l).

Table 13 lists the top-20 distinctive categories thus obtained for the two classes of users: *business* and *non-business*. We observe that there are significant and meaningful differences between the two user groups. Particularly, for businesses, we notice the presence of a large number of *service-oriented* categories such as "automotive services", "home renovation services", "computer services" and "cell phone services" and other *business-oriented* categories including "heavy equipment", "business/industrial" and "cargo trailers".

It can also be observed that the majority of distinctive categories for business class are actually sub-categories of *cars & vehicles* whereas for non-business users *buy and sell* makes up the major category. This is due to an overwhelming presence of non-business users in *buy and sell* category. Due to this fact, the only such categories which are determined as most distinctive for business users are the ones that are inherently service-oriented or business-oriented in nature (as mentioned previously) or the ones where the ratio of business to non-business users and their corresponding ads is not as imbalanced as in some of the others. Accordingly, we observe that "tickets" and "beds/mattresses" (furniture) categories are peculiar for businesses while some of the other furniture categories such as "couches/futons" and "chairs/recliners" are ranked as the top distinctive ones for non-business class.

5.3. *Temporal Changes in the Categories for User Groups*

The set of distinctive categories obtained for each user class in the previous section give a static picture of each class; in particular, it does not tell us which categories are popular with the respective user class at certain times and how those categories change over time. This is the object of our analysis in this section.

Let us denote with U_l the set of users who belong to a class l, and let $P(c \mid U_l)$ be the marginal probability of an ad category c over the entire duration of the dataset and $P_x(c \mid U_l)$ be the same probability over a time interval x. These probabilities can be derived as discussed in Section 5.2. An importance score of category c at interval x can then be computed as

$$I_x(c,U_l) = P_x(c \mid U_l) \ln \frac{P_x(c \mid U_l)}{P(c \mid U_l)} \tag{3}$$

Table 13. Distinctive categories for user groups.

S. No.	Categories
Business	
1.	(buy and sell, tickets)
2.	(cars & vehicles, cars & trucks)
3.	(buy and sell, business/industrial)
4.	(cars & vehicles, RVs/campers/trailers, cargo/utility trailers)
5.	(cars & vehicles, ATVs/snowmobiles, ATV parts/trailers/accessories)
6.	(buy and sell, home renovation materials, cabinets/countertops)
7.	(buy and sell, phones, cell phone services)
8.	(cars & vehicles, RVs/campers/trailers, RVs/motorhomes)
9.	(cars & vehicles, heavy equipment, other)
10.	(buy and sell, furniture, beds/mattresses)
11.	(cars & vehicles, motorcycles, motorcycle parts/accessories)
12.	(buy and sell, computer accessories, services (training/repair))
13.	(cars & vehicles, automotive services, towing/scrap removal)
14.	(cars & vehicles, automotive services, repairs/maintenance)
15.	(cars & vehicles, heavy equipment, heavy equipment)
16.	(buy and sell, computers)
17.	(buy and sell, home renovation materials, floors/walls)
18.	(cars & vehicles, automotive services, detailing/cleaning)
19.	(buy and sell, computers, laptops)
20.	(cars & vehicles, ATVs/snowmobiles, snowmobiles parts/trailers/accessories)
Non-Business	
1.	(cars & vehicles, auto parts/tires, tires/rims)
2.	(buy and sell, books)
3.	(buy and sell, electronics)
4.	(buy and sell, art/collectibles)
5.	(buy and sell, toys/games)
6.	(buy and sell, other)
7.	(buy and sell, phones/tables)
8.	(buy and sell, phones, cell phones)
9.	(buy and sell, jewellery/watches)
10.	(buy and sell, clothing, women's - tops/outerwear)
11.	(buy and sell, home - indoor, home decor/accents)
12.	(cars & vehicles, auto parts/tires, other parts/accessories)
13.	(buy and sell, hobbies/crafts)
14.	(buy and sell, furniture, couches/futons)
15.	(buy and sell, clothing, men's)
16.	(buy and sell, furniture, chairs/recliners)
17.	(buy and sell, sporting goods/exercise, exercise equipment)
18.	(buy and sell, tools, power tools)
19.	(buy and sell, baby items, strollers/carriers/car seats)
20.	(buy and sell, cameras/camcorders)

giving the degree at which postings in c for a particular user group U_l differ for interval x. Note that the probabilities are all conditional on the user group U_l being studied and a sudden drop in the number of ads for one user group in a category will not necessarily boost the importance of the category for other user groups.

We apply the aforementioned model to determine the most popular categories for each month covered by our dataset (May 2013 - January 2014) as compared to its entire time span for both our user groups. Table 14 shows the top five categories

Table 14. Distinctive categories for user groups over time.

Categories	**M**	**J**	**J**	**A**	**S**	**O**	**N**	**D**	**J**
Business									
(bs, art/collectibles)	-	5	2	2	2	3	2	2	2
(bs, computers)	4	2	4	3	5	-	-	-	-
(bs, computers, desktop computers)	-	-	-	-	-	-	-	-	5
(bs, computers, laptops)	-	-	-	-	-	-	4	4	4
(bs, furniture, beds/mattresses)	-	-	-	4	-	-	5	3	-
(bs, phones, cell phones)	-	-	-	-	-	4	-	-	-
(bs, phones/PDAs/iPods)	1	-	-	-	-	-	-	-	-
(bs, phones/tablets)	2	1	1	1	-	-	-	-	-
(bs, tickets)	-	-	-	-	1	1	1	1	1
(cv, ATVs/snowmobiles, ATVs parts/trailers/accessories)	-	3	-	5	4	-	-	-	-
(cv, RVs/campers/trailers, RVs/motorhomes)	-	-	3	-	-	-	-	-	-
(cv, RVs/campers/trailers, cargo/utility trailers)	-	4	-	-	3	5	-	-	-
(cv, RVs/campers/trailers, travel trailers/campers)	3	-	5	-	-	-	-	-	-
(cv, auto parts/tires, other parts/accessories)	-	-	-	-	-	-	-	5	3
(cv, auto parts/tires, tires/rims)	-	-	-	-	-	2	3	-	-
(cv, boats/watercraft, powerboats/motorboats)	5	-	-	-	-	-	-	-	-
Non-Business									
(bs, books)	-	-	-	2	1	-	-	-	1
(bs, clothing, costumes)	-	-	-	-	-	3	-	-	-
(bs, clothing, kids/youth)	-	-	-	-	5	-	-	-	-
(bs, clothing, women's - tops/outerwear)	-	-	-	-	-	5	3	-	-
(bs, computer accessories)	-	5	3	-	3	-	-	-	-
(bs, computers)	5	4	2	4	4	-	-	-	-
(bs, computers, laptops)	-	-	-	-	-	-	-	-	5
(bs, electronics)	-	-	-	-	-	-	-	4	-
(bs, garage sales)	-	3	4	3	-	-	-	-	-
(bs, jewellery/watches)	-	-	-	-	-	-	-	3	-
(bs, phones, cell phones)	-	-	-	-	2	1	1	1	2
(bs, phones/PDAs/iPods)	1	-	-	-	-	-	-	-	-
(bs, phones/tablets)	2	1	1	1	-	-	-	-	-
(bs, sporting goods/exercise, golf)	4	-	-	-	-	-	-	-	-
(bs, sporting goods/exercise, snowboard)	-	-	-	-	-	-	5	-	-
(bs, tickets)	-	-	-	-	-	4	4	2	4
(bs, toys/games)	-	-	-	5	-	-	-	-	-
(cv, ATVs/snowmobiles, snowmobiles)	-	-	-	-	-	-	-	5	3
(cv, RVs/campers/trailers, travel trailers/campers)	3	2	5	-	-	-	-	-	-
(cv, auto parts/tires, tires/rims)	-	-	-	-	-	2	2	-	-

Note: Column headers M-J represent months from May 2013 to January 2014. "bs" and "cv" denote *buy and sell* and *cars & vehicles* respectively.

for each month, projected over the categories shown.

The results reveal interesting trends. Specifically, for non-business users, we observe that books category experiences a strong surge during the months of August, September and January. This coincides with the beginning and ending of Fall term in the universities Likewise, costumes category sees a higher than usual traction during the month of October, which concurs with the Halloween season. This is reinforced by the fact that some other clothing categories became popular with the

non-business user group around the same period.

In the same way, some seasonal trends can be identified. For example, garage sale category experiences great activity at the onset of the summer season in Edmonton; electronics, jewellery and watches become especially popular during Christmas and start to lose traction afterward; tires/rims category witnesses abnormally high number of postings during October and November which marks the beginning of winter in Edmonton, thus indicating people eager to change the tires of their vehicles to withstand the harsh winter season.

Similarly, sporting goods, exercise and recreational categories also reveal interesting seasonal trends. We observe that the category pertaining to golf encounters a surge during summer while the one related to snowboards is popular at the arrival of winter. In addition, ticket category becomes active in winter months most probably due to interest in Edmonton Oilers games. We can also note that travel trailers, campers, RVs attract a lot of attention during summer while snowmobiles become popular with the non-business users during winter.

Many of the trends described above can also be noticed for the business users. For example, licensed ticket sellers are most active during the Oilers hockey season; automobile technicians and businesses post a large number of ads in tires/rims category at the commencement of winter season offering their services; activity in RVs, motorhomes, travel trailers, campers and boats categories sees an upward jump during summer.

At the same time, we can also observe the effect of discontinued or newly introduced categories over time. For example, it can be noted that *(buy and sell, phones/tablets)* category shows up in the list for both user groups initially. However, around September, it disappears from the ranking and never emerges later. The reason for this behavior is that it was discontinued during this time. Since users were not allowed to post in this category anymore, it gives the false impression that as compared to the entire duration of the dataset, the category under question was extremely popular with the users in the beginning. This is also true for *(buy and sell, phones, cell phones)* category which was introduced in place of the previously mentioned one and appears to experience huge activity in the later months.

5.4. *Distinctive Categories for Locations*

Our goal, in this Section, is to identify some of the locality patterns for users and their posts. In other words, we seek to find the categories that are the most distinctive or unusual for a particular location.

Following the approach taken by Backstrom et al.[?] for search engine queries, we model the ads posted in various categories as Bernoulli trials; the trial is a success if an ad is posted in a particular category c and a failure otherwise. Let p be the probability of success of the trial, computed as the fraction of overall ads posted in c, and t_x be the total number of ads posted from a specific location x. These quantities represent our binomial experiment, consisting of t_x trials, each with a probability

of success p. Considering that the individual trials are statistically independent i.e. the posting of ads in categories is independent of each other, the probability of s_x successes (i.e. the probability of seeing that many posts in category c from location x) is given by:

$$P(X = s_x) = \binom{t_x}{s_x} p^{s_x} (1-p)^{t_x - s_x}$$

A low probability value for a location x and category c indicates that the observed frequency is less likely according to the model, or it significantly differs from the global background distribution of c, making c a distinctive category for location x.

For the experiment, we used all categories in Kijiji instead of only the two top-level categories *buy and sell* and *cars & vehicles* used in our previous experiments. We divided the city into 38 neighborhoods (as shown in Figure 6), with each neighborhood starting with a specific 3-digit postal code, as issued by Canada Post. To obtain the appropriate neighborhood for each ad, we utilized the *address* attribute in each ad; ads not having postal code (nearly 24%) were ignored.

Table 15. Distinctive categories for various neighborhoods.

Categories	**T5H**	**T5J**	**T5K**	**T5S**	**T5V**	**T6C**	**T6G**	**T6P**	**T6S**
(bs, books)	-	-	3	-	-	-	2	-	-
(bs, electronics)	3	-	5	-	-	5	-	-	-
(bs, furniture, beds/mattresses)	-	-	-	-	3	-	-	-	-
(bs, furniture, dining tables and sets)	-	-	-	-	5	-	-	-	-
(bs, phones/tablets)	-	-	-	-	-	-	3	-	-
(bs, phones, cell phones)	-	-	-	-	-	-	1	-	-
(bs, sporting goods/exercise)	-	-	-	-	-	4	-	-	-
(bs, tickets)	5	5	1	-	-	3	-	-	-
(cv, cars & trucks)	1	-	2	1	2	1	5	-	5
(cv, auto parts/tires, body parts)	-	-	-	-	-	-	-	2	-
(cv, auto parts/tires, tires/rims)	-	-	-	3	-	-	-	-	3
(re, apartments/condos, 1 bedroom)	2	-	4	-	-	-	-	-	-
(re, apartments/condos, 2 bedroom)	4	-	-	-	-	-	-	-	-
(re, house rental)	-	-	-	-	-	2	-	-	-
(re, houses for sale)	-	-	-	-	-	-	-	3	-
(re, room rental, roommates)	-	-	-	-	-	-	4	-	-
(jobs, bar/food/hospitality)	-	3	-	-	-	-	-	-	-
(jobs, construction/trades)	-	-	-	2	1	-	-	1	1
(jobs, customer service)	-	4	-	-	-	-	-	-	-
(jobs, driver/security)	-	-	-	5	-	-	-	5	4
(jobs, general labour)	-	-	-	4	4	-	-	4	2
(jobs, office mgr/receptionist)	-	2	-	-	-	-	-	-	-
(jobs, sales/retail sales)	-	1	-	-	-	-	-	-	-

Note: "bs", "cv" and "re" stand for *buy and sell*, *cars & vehicles* and *real estate* respectively.

Table 15 shows the top five categories for some neighborhoods. Evaluating such qualitative results is a challenging task, however, we present some of our observations here. First of all, the distinctive categories in neighborhood T6G are the ones

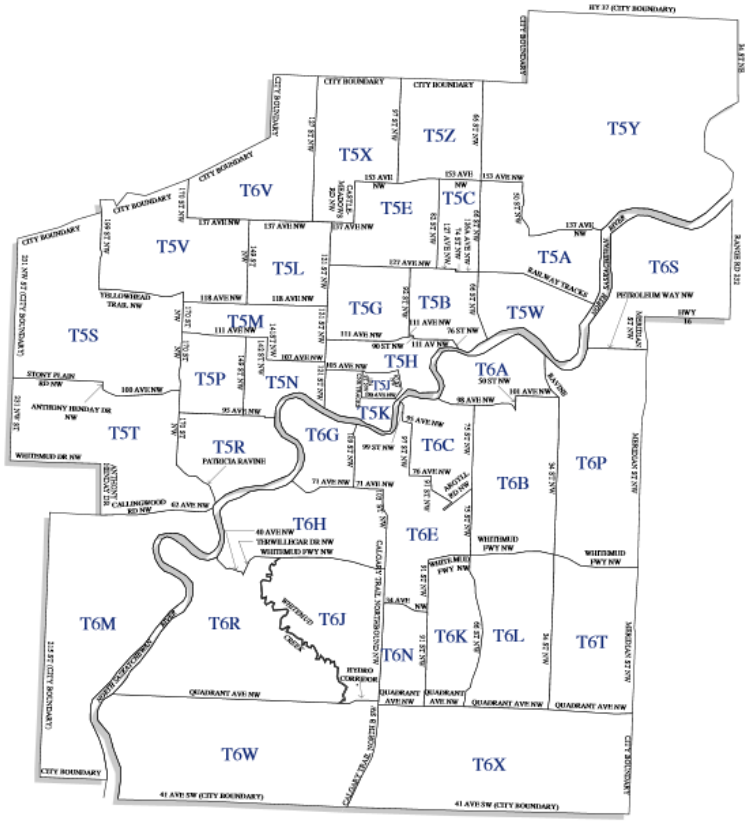


Fig. 6. Edmonton's Postal Code Map (© Canada Post, 2001).

in which we would expect great activity by the students (for example phones, books, tablets and room rental). This is not surprising since due to presence of University of Alberta here, the area is inhabited by many students. Similarly, books category is also popular in the nearby T5K locality due to the presence of MacEwan University in T5J neighborhood.

Moreover, we observe that T5J area experiences an unusually high number of ads from jobs categories. Again, this is to be expected since the area represents Edmonton Downtown, the hub of city's business activities. Due to this reason, as it can be easily imagined, there is always a high demand for accommodations around this neighborhood. Appropriately, we find that the nearby T5H, T5K and T6C areas attract a lot of postings from aparments/condos and house rental categories.

Finally, we observe that the nature of jobs required in Edmonton's downtown area (T5J), which is also the center of the city, is office-oriented i.e., desk jobs. On the contrary, as we move towards the outskirts of the city (T5S, T5V, T6P, T6S), we notice that the jobs become more physically demanding showing construction

and new housing activities.

6. Conclusions

In this paper, we sought to characterize users in a classified ad network and to use the results thus obtained to further analyze the network. Specifically, we studied the problem of determining if a user is utilizing a classified ad network for promotion of his/her 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 such business users. We experimented with various content-based features and classification algorithms and analyzed their performance in comparison to various baselines. We also validated our content-based ad classification approach on data obtained from another popular online classified ad network, Craigslist. 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 using the classified ad network.

In addition, we developed a probabilistic model of user behavior based on the ads the user posts and the categories in which the ads are posted. The model can track some of the temporal changes in behavior, as revealed by our experiments on two classes of users monitored over a period of almost a year. We studied the association between post categories and user groups, and showed how temporal and seasonal changes can be detected. Finally, we investigated a generative model for ad posts, based on user locations, and provide some evidence showing that the model is promising and that some interesting relationships can be identified.

This work can be extended in numerous directions. First of all, for nearly all the experiments, we used a subset of categories from Kijiji. A direction for future research can be to utilize the data from other categories for user classification, some of which (for example *jobs* and *services*) may not have the same degree of imbalanced data as was prevalent in the categories we studied. Another interesting direction is activity recognition and more complex patterns describing user activities. For example, a *downsizing* may be detected by a sharp increase in selling non-essential items. Similarly, an important avenue is to apply entity resolution to various aspects of classified ad networks, for example to determine if multiple ads are selling the same item or if the items are in the same condition. Finally, as a further application of this research, many more aspects of the classified ad network can be analyzed. Predicting how users determine the price of an item when listing an ad and identifying what makes a particular classified ad get more views than others are just two such examples.

Acknowledgments

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