

Characterizing Users in a Classified Ad Network

by

Muhammad Waqar

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

© Muhammad Waqar, 2014

Abstract

We study the problem of classifying users in a classified ad network and its applications in further analyzing the network. Specifically, we seek to classify Kijiji users into one of the two *business* and *non-business* categories. The problem is challenging due to the sparsity of the data about users, the vague separation of the two classes, and the highly imbalanced distribution of users between the classes. Our work utilizes the ad content to build a set of distinctive terms for each class (profile). Given the statistics on how an ad mentions terms from a class profile, the affinity of an ad (and subsequently a user) to a particular class is determined. Our experiments reveal that this is an effective strategy for classifying users, outperforming various baselines. We study the impact of profile size on the classification task and observe that using longer class profiles may not be helpful. Moreover, 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 also study the same problem from a different angle: collective behavior of a user in posting ads. Using features associated with such behavior, we identify four distinct usage patterns for the users of the Kijiji network and study the association of business and non-business users with these patterns. Our experiments reveal that a sizeable number of members from both user groups validly manifest all the patterns, due to which the aforementioned features are inadequate for the classification task.

Finally, using the results of user classification, we analyze the Kijiji network from various aspects. Our results, for example, indicate that businesses are more amenable to post consistently in a particular set of categories than non-business users and that the popularity of different categories for both the user groups exhibits various seasonal trends.

Acknowledgements

First and foremost, I thank the Almighty for providing me with the opportunities to learn, progress and succeed throughout my entire life.

This work would not have been even remotely possible without the support and guidance of my supervisor, Dr. Davood Rafiei. Thank you for your availability, sound advice, patience, perseverance, financial assistance, encouragement, mentorship, tireless revision, insightful comments, and most importantly, the opportunity to work on some interesting projects.

I am also grateful to Dr. Denilson Barbosa, without whose motivation and encouragement I would not have pursued graduate studies in Databases. I would also like to thank Dr. Sarah Moore from the Alberta School of Business for taking time out from her busy schedule to serve as an examiner.

I must also acknowledge my office mate and friend, Jeeva Paudel, for all the wisdom, camaraderie and advice he provided throughout the course of my graduate studies. I am also deeply indebted to my friends, Nael, Usman and Aryn, for our debates and helping me settle in a foreign country.

Special thanks goes out to Aibek Makazhanov and Afsaneh Esteki for their assistance with the research.

I would also like to express my gratitude to my brother, Shahab, and sister, Hina, for their love and support at all times.

Lastly, I wish to thank my parents, Shehnaz Hameed and Abdul Hameed, who never compromised on the quality of education for their children, and whose unconditional love and prayers remain my unrelenting source of strength. To them, I dedicate this thesis.

Table of Contents

1	Introduction	1
1.1	Thesis Statement	3
1.2	Research Contributions	3
1.3	Thesis Organization	4
2	Related Work	5
2.1	Text Classification	5
2.2	User Modeling in Social Media	6
2.3	Social Network Analysis	7
3	Data Collection	9
3.1	Crawling Edmonton Kijiji	10
3.2	Dataset Statistics	12
3.3	Visualization	13
4	Problem Definition and Experimental Setup	14
4.1	Problem Formulation	14
4.2	Experimental Setup	15
4.2.1	Dataset Preparation	15
4.2.2	Classifiers	17
4.2.3	Dealing with Imbalanced Data	17
4.2.4	Evaluation Metrics	18
5	Using Content to Identify Business Users	20
5.1	Motivation	20
5.2	Strategy	21
5.3	Building Profiles	21
5.4	Methodology	25
5.4.1	Profile Features	26
5.4.2	Non-Profile Features	26
5.5	Dataset Preparation	28
5.6	Results	29
5.6.1	Ad Classification	29
5.6.2	Impact of Profile Size	31
5.6.3	Feature Analysis	33
5.7	Classifying users	35
5.7.1	Experimental Evaluation	37
5.8	Using Unlabeled Data	39
5.8.1	Experimental Evaluation	40

6	Studying Users by Posting Behavior	42
6.1	Motivation	42
6.2	Behavioral Features	43
6.3	Experiments and Evaluation	47
6.3.1	Classifying Users Based on Posting Behavior	47
6.3.2	Usage Patterns	48
7	Analysis of the Classified Ad Network	53
7.1	Temporal Changes in User Profiles	54
7.2	Distinctive Categories for User Groups	56
7.3	Temporal Changes in Popular Categories for User Groups	58
7.4	Distinctive Categories for Locations	63
8	Conclusions	68
8.1	Future Work	69
	Bibliography	71

List of Tables

3.1	Statistics for complete and abridged datasets.	13
4.1	Statistics for dataset prepared for user classification task.	17
4.2	Confusion matrix for a binary classification problem.	18
5.1	Top ranked bigrams from the class profiles.	25
5.2	Sample calculation of <i>number of mentions</i> feature over different domains.	27
5.3	Statistics for dataset prepared for ad classification task.	29
5.4	Results for ad classification. Profile size is set to 100.	30
5.5	Feature ranking for ads classification task.	33
5.6	User classification results. Profile size is set to 200.	38
5.7	Results for users classification using unlabeled ads dataset. Profile size is set to 200.	40
6.1	Statistics for filtered dataset (users who have posted at least two ads).	47
6.2	Ranking of posting behavior features.	48
6.3	Results for users classification.	48
6.4	Confusion matrix of the clusters automatically detected by EM algorithm along with the manually annotated data and the percentage of a particular class of users in each cluster. Features used for clustering are listed in Table 6.2.	49
6.5	Mean value of features for clusters (Table 6.4). Standard deviations are shown inside brackets. Feature identifiers are listed in Table 6.2.	50
7.1	Distinctive categories for user groups.	57
7.2	Popular categories for non-business users over time. Column headers M-J represent months from May 2013 to January 2014. “bs” and “cv” denote <i>buy and sell</i> and <i>cars & vehicles</i> respectively.	60
7.3	Popular categories for business users over time. Column headers M-J represent months from May 2013 to January 2014. “bs” and “cv” denote <i>buy and sell</i> and <i>cars & vehicles</i> respectively.	61
7.4	(1/2) Distinctive categories for various neighborhoods. “bs”, “cv” and “re” stand for <i>buy and sell</i> , <i>cars & vehicles</i> and <i>real estate</i> respectively.	66
7.4	(2/2) Distinctive categories for various neighborhoods. “bs”, “cv” and “re” stand for <i>buy and sell</i> , <i>cars & vehicles</i> and <i>real estate</i> respectively.	67

List of Figures

3.1	Schematic view of the network data.	10
3.2	Data crawled from an ad posted on Edmonton Kijiji network.	11
3.3	Distribution of users by their posted ads.	12
3.4	Distribution of ads in various categories.	12
4.1	Example of an ad posted by a business for other (private) purposes. . .	15
5.1	Sample Business and Non-Business ads.	22
5.2	Impact of varying profile size on results of ad classification.	32
5.3	Distribution of training examples across different feature spaces. . .	34
5.4	CDF of fraction of business ads detected by our classifier against the percentage of users in our dataset.	36
6.1	Sample Business and Non-Business users.	44
7.1	Temporal Changes in User Profiles.	55
7.2	Edmonton's Postal Code Map. © Canada Post, 2001.	64

Chapter 1

Introduction

With a tremendous growth of the World Wide Web (WWW) in the past few decades, people have been utilizing this medium to address all aspects of their lives, both at home and at work. As a consequence, many businesses and industries have made or are making a switch to the new medium. Classified advertising is one such example where greater convenience and cheaper costs are driving more and more individuals away from traditional print ads and more towards online classified advertising. Although these ads are mostly placed by private individuals to sell or buy a particular item, many businesses are also using this medium for the promotion of their products and/or services, finding the right job applicants, etc. This is due to the fact that classified ads now have a relatively large user base, and are much more inexpensive than TV/radio commercials or billboard advertising traditionally used by businesses.

It is important to differentiate between classified ad networks on the web and e-commerce retail sites such as Amazon and AliExpress. The former provides a way to list items, services, community events or properties for sale often for free 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, places a greater emphasis on the satisfaction of their users by increasing their quality and reliability as well as protecting users from scams. Therefore, they often charge a fee for listing an ad, allow users 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 these features, distinguishing between the two

user groups in a classified ad network becomes increasingly difficult.

Ascertaining if a user in a classified ad network is a business or a private individual involves various other 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 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 users better information about the nature of the seller. Also, the government may need to identify businesses or gather some information about them for different purposes such as taxing. The online network itself may use this information to analyze its pricing strategy and other potential sources of revenue. Furthermore, such an automatic identification of users can be helpful in automating the process of creating web directories which takes considerable time and effort if humans were to discern the businesses manually. Moreover, such data can help in better understanding of the dynamics of the classified ad networks.

In this work, we treat the problem as a binary classification task where given a user and his posted ads, the goal is to detect if the user is a *business* (using the network for promotion of his enterprise) or *non-business* (private individual) user. To the best of our knowledge, this is the first work that studies such a classification. Our approach 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. We study the issues related to weighting of the terms and the effects of varying profile size on the classification results. In addition, we propose a simple bootstrapping heuristic in cases when labeled data is not available for a supervised classification. We also investigate the problem in the context of a different set of features, based on the posting behavior of the users and some of the

prevalent usage patterns.

Postings of a classified ad network may have more structured attributes such as the category, time and location (address) associated to the listing etc. We explore some of the relationships between these attributes to gain a better understanding of the network. For instance, we study how amenable user groups are to post consistently in a specific set of categories over time. Likewise, we also analyze the temporal changes in popular categories for users as well as the distinctive categories for various neighborhoods. The conclusions that we draw and trends that we observe here can have many practical applications. For example, the information corresponding to which categories become popular at a particular time of year can be leveraged by the classified ad network to announce various posting deals for business users during a specific period.

1.1 Thesis Statement

Our thesis statement is that the content of a user's posted ads as well as his posting behavior in an online classified ad network is effective for determining if the user is utilizing the network for the advocacy of his enterprise or for personal needs. Also, such a classification can aid us in understanding the dynamics of the respective user groups within the network and reveal more insightful patterns.

1.2 Research Contributions

The following are the contributions of this thesis:

- A study on classifying users of a classified ad network into *business* and *non-business* classes based on the content of the ads. Specifically, a classifier for distinguishing between the two user groups based on the text of user postings and an experimental evaluation showing its performance and the effectiveness of the features studied.
- A study of the aforementioned user classification task using the posting behavior of the users. Particularly, a classifier built on collective behavioral

features of the users in posting ads and an experimental evaluation showing its performance.

- Analysis of the classified ad network from various aspects, specifically determining temporal changes in user profiles, popular categories for user groups over time and distinctive categories for user groups as well as neighborhoods.

1.3 Thesis Organization

The rest of this thesis is organized as follows. In Chapter 2, we review the related literature and compare its similarities and differences to our work. In Chapter 3, we describe the process by which dataset used in this work was collected. In Chapter 4, we present our problem formulation and describe the general experimental setup for the next two chapters. Chapter 5 details our methodology for classifying users using ads content and corresponding results in a supervised as well as semi-supervised setting. Chapter 6 presents the second part of the study, which is the analysis of the same user classification problem using posting behavior of the users and the results of a classifier using such features only. In Chapter 7, we analyze the network from various dimensions using the results of user characterization obtained previously. Finally, in Chapter 8, we summarize our conclusions and present various avenues for future research.

Chapter 2

Related Work

Classified ad networks have not received much attention from the research community (particularly from the computer science community) as compared to some of the other networks. Some studies have been reported on Craigslist, studying its impact on local newspapers [29, 43] and linking it to the spread of sexually transmitted diseases [10]. Furthermore, researchers have investigated the usage of sexual health-related language in classified ads [22] as well as the movement behavior of anonymous, casual sex-seeking individuals [19] in Craigslist network. However, we are not aware of any studies on characterizing users of a classified ad network or a quantitative analysis of their behavior on a large scale. That said, our work is related to the body of work on text classification, user modeling in social media and social network analysis.

2.1 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 was focused on categorizing documents by topics. Several techniques have been developed for this purpose, and a relatively comprehensive survey of them can be found in [1, 8]. A major challenge in text classification has been the high dimensionality of the feature space, since the native feature space consists of unique terms in a document which can easily be tens or hundreds of thousands in number for a moderate-sized text collection. Yang and Pederson [48]

evaluate five methods for feature selection and find information gain, χ^2 -statistic and document frequency to be most effective for ranking features. Examples of topical text classification can be found in spam email detection [41], classifying news stories [14] and blog posts [47], etc.

More recently, there has been a growing interest in the field of non-topical classification, which is more related to our work. For example, Mishne [32] classifies blog posts based on the mood of the writers; Eickhoff et al. [16] identify if web pages are suitable for children or not and Pierre [40] performs a classification of websites into industrial categories. There has also been related work on detecting the online commercial intent of the users based on their search queries and visited web pages [15, 24]. A more closely related work to ours is that of Makazhanov et al. [31] who use interactions with a party to determine the political preference of Twitter users. However, unlike them, 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.

2.2 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. [30] 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. [2] study the same problem in the context of the Twitter network, utilizing tweets posted by users to infer their preferences. Schöfegger et al. [42] analyze the tagging behavior of users in a social academic network to predict their research discipline. Stoyanovich et al. [46] also investigate the tagging behavior of users and 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 comprehen-

sive survey of such systems and approaches can be found in [3].

Benevenuto et al. [6] 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 [12], commercial campaigns in Community Question and Answer (CQA) forums [11] and spammers as well as content promoters in online video social networks [7].

2.3 Social Network Analysis

Online classified ad networks 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 relevant.

A great deal of work has been done to gain insights into the users' behavior in online social networks (OSNs). For example, Guo et al. [21] analyze three OSN workloads and report that users' posting behavior exhibited strong daily and weekly patterns. Similar studies have also been reported on microblogging networks [20] and on web search engines [25, 26, 45] to explore their usage by people and the searching trends that emerge over time. Beitzel et al. [5] 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 network from various dimensions in order to gain a better understanding of it.

Related work also includes the literature on finding groups of users or *communities* whose members share a similar profile, i.e., exhibit a similar behavior in their interaction with a system [35]. The proposed techniques include conceptual clustering [18], cluster mining [39], modularity based approaches [13, 33], graph

partitioning approaches [27, 34], clique percolation [38], etc. Some of these techniques may also be applicable in the setting of a classified ad network, to find user groupings with similar interests.

Chapter 3

Data Collection

For our work, we collect advertisements from Kijiji ¹, a popular online classified ad service that allows users to post free classified ads in different categories. It is organized around local communities and is structured as a network of sub-websites (sites). Each site contains ads from its primary anchor city, as well as from smaller surrounding communities. It is a subsidiary of eBay and was launched in March 2005. Kijiji has presence in more than 300 cities in Canada, Italy, Hong Kong, Taiwan, Switzerland, Turkey, India, Austria and United States (where it was re-branded as eBay Classifieds).

We chose Kijiji for our work because of its popularity in Canada and the fact that it allows users to register an account with the website. As such, each user account is associated with a unique identifier. This id allows us to connect each ad to the user who posted it, 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 ² (possibly with minor modifications) as long as users can be identified.

Figure 3.1 depicts a schematic view of the network data. Users are allowed to post ads at any point in time, many of which may be active simultaneously. Each ad belongs to a particular category and is generally (though not always) associated with a location.

¹<http://www.kijiji.ca>

²<http://www.craigslist.org>

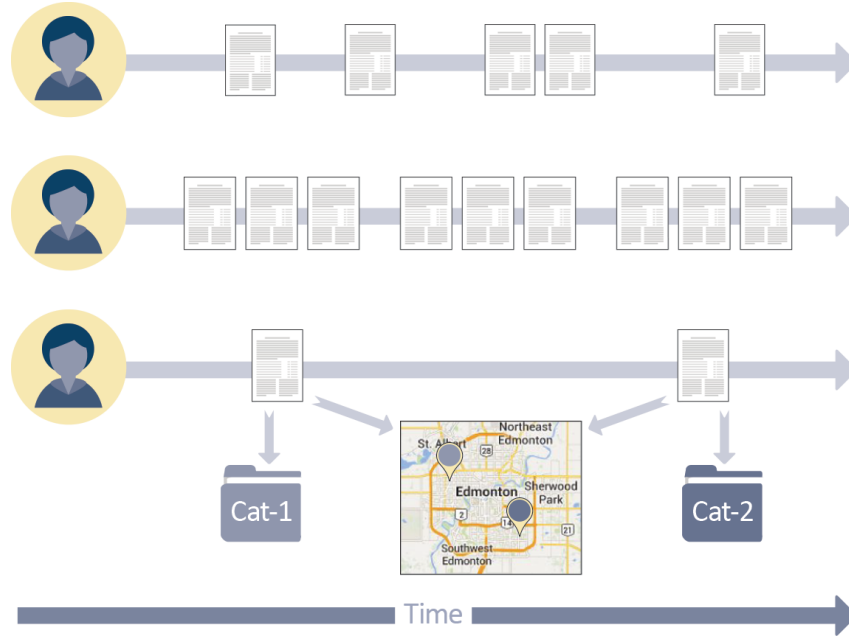


Figure 3.1: Schematic view of the network data.

3.1 Crawling Edmonton Kijiji

We built a crawler to extract the ads from Edmonton Kijiji ³ which services the cities of Edmonton and St. Albert as well as the nearby Strathcona County. During each crawling session, our crawler went through the ads, newly posted since the last session, active at the time and extracted their various fields. The ads previously detected in the database were ignored. Specifically, for each ad, we extracted the following (refer to Figure 3.2 for the corresponding locations on the site):

1. **ID:** A unique integer assigned to each ad posted on Kijiji.
2. **Title:** A short description of product (item) or service being offered (or required, in which case the title starts with *Wanted*).
3. **Category:** Ads are organized into various categories on Kijiji, each of which may have many sub-categories. Each ad category is a 6- or 7-tuple. Some members of the tuple are reserved for location information pertaining to the particular sub-website of Kijiji being used, which is not useful for our work. Thus, before any experiment, we pre-process the category field to remove

³<http://edmonton.kijiji.ca>

The screenshot shows a Kijiji advertisement for a queen bed. The page layout includes a header with the Kijiji logo, navigation links (Sign In, Register, Help, Français), and a search bar. The ad is titled "Brand NEW Queen Bed! Call 780-437-0808!". The ad details include: Date Listed: 11-Jul-14, Price: \$300.00, Address: 6408 Gateway Boulevard, Edmonton, AB T6H 2H9. The ad is for sale by a dealer. The description states: "Brand New, Still in the original factory wrap! Queen bed includes: Headboard, Footboard & Rails! Other pieces may be available. Delivery is available! Call today! 780-437-0808!". The ad has 341 visits. The poster contact information includes a phone number 780-437-XXXX and a link to view other ads. The email poster section includes a form for email, name, and message, with a verification code 9597.

Figure 3.2: Data crawled from an ad posted on Edmonton Kijiji network.

such information so that the category (*Alberta, Edmonton Area, Edmonton, buy and sell, furniture, beds/mattresses in Edmonton*) becomes (*buy and sell, furniture, beds/mattresses*). Throughout this work, we adopt the notion that the category tuple (*buy and sell, bikes*) is inclusive of all the sub-categories inside it.

4. **Attributes:** Various attributes of the ad. The only attribute guaranteed to be present in every ad is *Date Listed*. Other attributes vary depending on the category in which the ad is posted. Some other popular attributes are *Address*, *Price*, etc.
5. **Description:** Details of the ad. We strip this HTML text and store it in a plain text format.
6. **User ID:** A unique integer assigned to the user posting the ad. This can be utilized to query other ads posted by the same user.

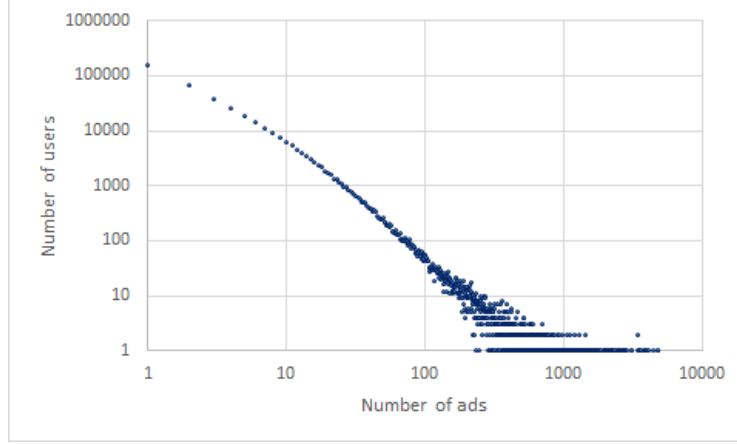


Figure 3.3: Distribution of users by their posted ads.

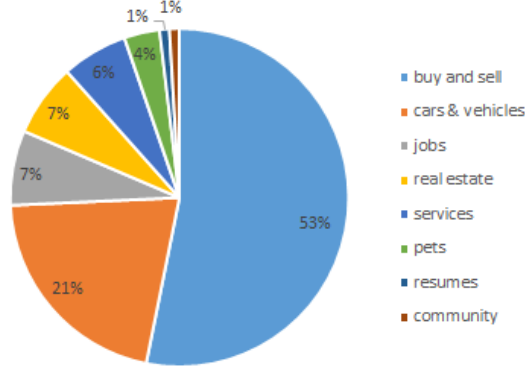


Figure 3.4: Distribution of ads in various categories.

3.2 Dataset Statistics

We ran the crawler once every day from May 1, 2013 to January 31, 2014. During this time, we were able to obtain millions of ads posted on the site. Figure 3.3 shows the log-log plot of the ads distribution for the users in our dataset. We observe that the ads distribution seems to follow a power law.

Figure 3.4 shows the distribution of crawled ads in different categories. An inspection of the list of sub-categories on the Edmonton Kijiji website shows that certain categories are very diverse in their extent, for example (*buy and sell*) allows users to purchase or offer for sale items ranging from books to entire businesses. On the other hand, many categories are relatively limited in their scope, such as (*pets*). For nearly all the experiments in this work, we utilize data from one representative of both the groups. Specifically, we selected (*buy and sell*) category from the diverse

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

Table 3.1: Statistics for complete and abridged datasets.

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 the latter group as it accounts for the most number of ads in the dataset after (*buy and sell*) (Figure 3.4). We refer to these two categories as the *abridged dataset*.

Table 3.1 reports some statistics for the complete as well as the abridged dataset.

3.3 Visualization

We built a system ⁴ for efficient browsing of the ads and users in the dataset. It can be used by all and sundry and was useful for preparing datasets for individual experiments (discussed in the respective chapters).

⁴<http://webdocs.cs.ualberta.ca/~mwaqar/kijiji/>

Chapter 4

Problem Definition and Experimental Setup

4.1 Problem Formulation

The main problem studied in this thesis can be formulated as follows:

Given a user u and the set of ads a_u that u has posted, predict if u is a *business* or a *non-business* user.

An important question from the problem formulation is how we characterize someone as a business user. For this work, we adopt the standard definition of a business from college textbooks, i.e. *a business is an organization involved in the trade of goods, services or both to consumers* [36]. They may be privately owned, not-for-profit or state-owned and may take any of the ownership form. However, it is to be noted that it is possible for business users to use the classified ad network for their personal use. This follows from the fact that not all enterprises use classified ads to advocate their offerings. An example of such user is shown in Figure 4.1. While it is clear from the ad description that the user owns an enterprise, one can easily recognize that the reason for this posting is to offer some tires at a reduced cost to free up some space in his shop. Thus, it is straightforward that within the network, such users act in accordance with the *modus operandi* of non-business users. In lieu of this, for our work, we require all businesses to be using the medium of classified ads to promote their products and/or services.

Visualizing Edmonton Kijiji

[Home](#) | [Browse User Ads](#) | [Browse Ads](#)

Ad ID 476636893

User ID

Z17579183 

Title

195/75R14 winter tires

Category

Kijiji Alberta | Edmonton Area | Edmonton | cars & vehicles | auto parts, tires | tires, rims

Description

I have a great set of studded winter tires that is taking up shop space. Hankook 14" tires. They are on black steel rims with 5 bolt pattern. Believe they fit a dodge k-car. Not sure though. Tread hardly looks worn they are in great shape and want them gone. Asking \$400 and open to offers. Email or text (780)608-4358. Thank you This ad was posted with the Kijiji mobile app .

Attributes

Date Listed: 20-Apr-13

Price: \$400.00

Last Edited: 24-Apr-13

Address: Camrose, AB, Canada

Figure 4.1: Example of an ad posted by a business for other (private) purposes.

A question then arises as to how to treat the users who use the network for both business and non-business purposes. We believe this is an interesting issue and convincing arguments can be made to place such users on either side. However, in this work, we consider these users as business users.

4.2 Experimental Setup

In this section, we describe the general setup of the experiments carried out in Chapters 5 and 6. We posit the problem as a binary classification task where given the set of ads posted by a user in a particular time interval (in our case the entire duration of the dataset), the goal is to predict one of the labels *business* or *non-business* for the user.

4.2.1 Dataset Preparation

As mentioned in Section 3.2, we use the abridged dataset in this work. To prepare the ground truth for our experiments, a random sample of the users in the dataset was manually labeled into *business* and *non-business* classes. The system described in Section 3.3 was used for this purpose. The annotators would enter the given list of user ids into the system which, in response, would display a summary view

(consisting of ad title, category and a brief description) of all the ads each user has posted. The annotators had the option of viewing each ad in detail. Since distinguishing between the two classes is sometimes confusing, annotators had the additional option to mark a user as *unknown*.

We realize that the manual labeling of users in our dataset is a very daunting task. This is due to the fact that a user may have posted multiple ads, and coming up with a class label for such a user requires a thorough investigation of all of these ads. What makes the task even more complex is that in some cases, no single ad provides enough evidence on its own to decide the classification of the user and the annotator has to aggregate various clues spread throughout multiple user's ads. Moreover, from the preliminary testing, we knew that the dataset is highly biased towards non-business users. This created a very precarious situation for us since we did not want to over-burden the annotators lest they tag the users clumsily, while at the same time, we needed to have sizeable number of users for both classes in our dataset for the experiments. We decided against having multiple judges tagging different parts of the data, since, we realized that each annotator has a different mental model of how businesses should manifest themselves in a classified ads network and we did not want to introduce inconsistency in data labeling.

In lieu of this, our first annotator tagged all the users in the random sample. Some of those tagged users were clear cut cases, meaning that the annotator was convinced that there were sufficient evidence to label the user; classifying such users was relatively easy. However, it was not the case for some of the others; even though these users were tagged, the first annotator was not fully convinced that the label was accurate. To reduce the workload of the annotators, from the clear cut cases, the users tagged as non-business and unknown were not passed to the second judge. More specifically, the data passed to the second judge consisted of (1) all the users tagged as businesses, and (2) all the users marked as requiring another set of eyes by the first judge. The latter set often included users who posted many ads, thereby, introducing some of the classification challenges described above. For our experiments, we selected clear-cut non-business users tagged by the first annotator, and those users for which both judges agreed upon and the label was either business

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

Table 4.1: Statistics for dataset prepared for user classification task.

or non-business (meaning that the unknown users were ignored). Annotators agreed on 70% of the users.

Some statistics of the dataset thus prepared are presented in Table 4.1.

4.2.2 Classifiers

We experimented with different classifiers provided in the Weka toolkit [23] and chose the ones that performed best in the preliminary tests, namely decision tree based Random Forest (RF), SVM based SMO and Logistic Regression (LR). We set parameters of the classifiers to their default values in Weka.

4.2.3 Dealing with Imbalanced Data

A dataset is said to present a class imbalance if it contains many more examples of one class than the other. Most machine learning methods suffer greatly when faced with severely imbalanced data since they are designed to optimize overall accuracy without taking into account the relative distribution of each class. As a result, these classifiers tend to ignore smaller classes while concentrating on classifying the large ones accurately. Unfortunately, this scenario is prevalent in many domains, for instance anomaly detection, fraud detection etc. as well as our dataset.

A large number of approaches have been proposed to deal with class imbalance problem. From [17], these approaches can be divided into two broad groups: *internal* approaches introduce algorithms or modify existing ones while *external* approaches use unmodified existing algorithms, but resample the data before feeding it to them to diminish the effect of the class imbalance. While the *internal* approaches can be very effective, they face a disadvantage of being application or domain specific. Estabrook et al. [17] explain this as a problem because datasets presenting different characteristics are better classified by different algorithms. Thus in our

		True Labels	
		Positive	Negative
Classifier Predictions	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 4.2: Confusion matrix for a binary classification problem.

work, we chose to utilize the *external* approaches to mitigate the unwanted effects of imbalanced data.

Specifically, we follow the recommendation by Klement et al. [28] and combine random under-sampling 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*). Finally, the individual classifiers are combined by averaging their predicted confidence.

For all the experiments in this work, we use 3 classifiers in the ensemble. Moreover, we also present results when the classifier is trained on the imbalanced data for comparison purposes.

4.2.4 Evaluation Metrics

We evaluate the quality of the classification in terms of F-measure, the harmonic mean of precision and recall, for every class label. Table 4.2 shows the confusion matrix for a binary problem which is used to compute the precision and recall of the classifier.

- **Precision (P):** Precision is defined as $\frac{TP}{TP+FP}$. In binary problems, precision describes the proportion of actual positive examples that are correctly identified.
- **Recall (R):** Recall is defined as $\frac{TP}{TP+FN}$. In binary problems, recall measures the fraction of positive examples that are correctly retrieved.
- **F-measure (F):** F-measure is a harmonic measurement defined as: $2 \times \frac{P \times R}{P+R}$.

It should be noted that for imbalanced datasets such as ours, accuracy $\frac{TP+TN}{TP+FP+TN+FN}$ is not a good metric. This is because the accuracy of the classifier can be high even

if it predicts dominant class label for every instance. Evaluating by F-measure for all classes, we avoid any such problem.

Chapter 5

Using Content to Identify Business Users

5.1 Motivation

There are a number of ways for approaching the user classification problem. In this chapter, we will study this problem by considering the content of the ads posted by the users. To motivate this approach, let us consider the two ads shown in Figure 5.1, and suppose that these are the only ads posted by their respective posters. Going through the content of the ad descriptions and having no other information about these users, one may easily tag the user who posted the ad in Figure 5.1a as a business user whereas the poster of the ad in Figure 5.1b is very likely to be labeled as non-business. By analyzing how so quickly we came about these decisions, it becomes apparent that the text of user postings provides important clues regarding the classification of the user. Specifically, only a few terms mentioned in the ads can make the distinction between two classes. For example, we do not expect to see 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 anticipate many business users to use the phrases *just upgraded*, *make me an offer*, etc. in their postings. Moreover, now assuming that both of these ads were posted by the same user (which is not implausible since as explained in Section 4.1, many business users also use classified ad network for their personal purposes and our definition of businesses caters to this behavior), one may still label the user as a business. Hence, we conclude that it may not be necessary for us to

examine all the ads of the user, rather in some cases, even a single ad can contain enough evidence to label the user as a business user. It is to be noted that the same does not hold for non-business class.

5.2 Strategy

In light of this motivation (Section 5.1), we study the problem of user classification as the classification of individual ads of the user. Specifically, in this chapter, we study the following problem:

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

Later in the chapter, we will use the results of ads classification to come up with a label for the user. Note that henceforth the phrase *ad classification* will be used to refer to the above problem.

Let us consider how each target class can be represented in the context of the ad classification problem. From the previous section, we realize that each n-gram has a particular relevance to each class. Hence, 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* (following from the fact that the profiles are ranked and weighted), we will seek to characterize an ad as being business oriented or non-business oriented. Note that it is possible for an ad to not mention any term from both the profiles. Such an ad may be tagged as *unknown*, meaning that it does not belong to either of the classes, consistent with our treatment of users in Section 4.2.1.

5.3 Building Profiles


In this section, we will describe the method for building profiles of the classes. For this purpose, we make use of the term-weighting scheme proposed in [44].

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

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

Figure 5.1: Sample Business and Non-Business ads.

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 $L = \{business, non-business\}$. 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 denoted as V .

We proceed to build the language model (LM) for each class as well as the entire corpus. The main idea behind this approach is to calculate Kullback-Liebler (KL) divergence between the LM probabilities of each class corpus and the entire corpus. The divergence score of an individual term can then be used as a measure of importance of the term to a specific class. In this way, we will be able to get a ranked and weighted list of class-specific terms.

In this model, $tf-idf$ scores are used to calculate term probabilities for a particular corpus. For the entire corpus, the marginal probability of a term is calculated as:

$$P(t \mid D) = \overline{tf}(t, D)udf(t, D)$$

and normalized as:

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

where $\overline{tf}(t, D)$ represents the average term frequency of t in the documents (ads) in D 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:

$$w(t \mid l) = \overline{tf}(t, D_l)udf(t, D_l)idf(t, D)$$

$$w^N(t \mid l) = \frac{w(t \mid l)}{\sum_{t \in V} w(t \mid l)}$$

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 under:

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

where the normalization factor λ is set to 0.001 as in [44].

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

$$P(t | l) = \frac{w^S(t | l)}{\sum_{t \in V} w^S(t | l)}$$

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

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

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:

$$I(t, l) = P(t | l) \ln \frac{P(t | l)}{P(t | D)} \quad (5.1)$$

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.

Now that we have a ranked and weighted list of terms for each class, the final step in building class profile is to choose the top terms from this list. This can be accomplished by selecting top-N terms from the profile or by selecting all the terms having importance score greater than some threshold. We experimented with both techniques in preliminary experiments and decided to use the former approach.

Table 5.1 lists some of the top-ranked bigrams from the class profiles of categories in the abridged dataset (Section 3.2). 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. Moreover, 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. [37]. On the other hand, non-business users are more likely to mention first-person singular pronouns (I, me, mine) in their ads.

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

Table 5.1: Top ranked bigrams from the class profiles.

5.4 Methodology

From Section 5.2, let us recall that we build profiles for each class, which is a set of terms that have a particular relevance to the said class. An ad may *mention* some or none of the terms from the profile of a class. In the latter case, it is not possible for us to determine the affinity of the ad with the relevant class.

To classify the ads, we employ a one-vs-all classification strategy where for each class label, we train a binary classifier. The classifier takes various features built on the statistics derived from the profiles (referred to as *profile* features) and some *non-profile* features (detailed in Section 5.4.2) and predicts the confidence with which an ad can be considered as having inclination towards the respective class. As we discussed earlier, since it is not always possible to reliably classify an ad at all, we use a standard threshold of 0.5; ads having 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. It should be noted though that unknown ads are not ignored in the user classification task, and their confidence scores, even if rather low, are used in aggregation when deciding on a class label for users (see Section 5.7 for details). We assume an ad cannot belong to both *business* and *non-business* classes. Therefore, if an ad mentions terms from both profiles and ends up with confidence scores above the threshold for both classes, then it is assigned to the class with the highest predicted confidence.

5.4.1 Profile Features

Given an ad, we use five statistics (features) based on how it mentions the terms from the class profiles. These are (1) number of mentions (2) average weight of mentions (3-5) average/min/max rank of mentions. Feature (1) is used because it is likely that the more the ad mentions terms from the profile of a particular class, the more tilted it would be towards that class. The intuition behind features (2-5) is that the higher a term is on the class profile, the more relevant and distinctive it is to that class. Hence, these features capture the relevance or importance of the ad mentions to a class.

Due to the fact that an ad may mention terms from both the classes, we calculate the profile features in the context of various domains. The idea of different feature domains was first introduced by Makazhanov et al. [31]. The notion of domains allows us to provide classifiers with *overall* statistics, calculated over all the classes whose profile terms are mentioned by the ad, and *relative* statistics, calculated for a particular class in relation to all the classes.

Let us illustrate the concept of domains using an example shown in Table 5.2. Suppose that an ad mentions terms from business profile 8 times and those from non-business profile 12 times. These are the values for *number of mentions* statistic calculated in the target or *T-domain* in the feature vectors for respective classifiers. In the overall *O-domain*, the same feature will be calculated as sum over all classes, and will have the value of 20 for both the feature vectors. In the relative or *R-domain*, the same feature is calculated as the fraction of its values in *T-* and *O-domains* i.e. $8/20 = 0.4$ for business class and $12/20 = 0.6$ for non-business. Finally, in the delta or Δ -*domain*, the feature value is the difference of the values in *T-* and *O-domains*. Thus, in all, each classifier uses 20 profile features.

5.4.2 Non-Profile Features

In addition to the statistics derived from profile mentions, we also use various other features that can provide useful information about the orientation of an ad. We divide them into three categories:

Domain	Business	Non-Business
T	8	12
O	20	20
R	0.4	0.6
Δ	-12	-8

Table 5.2: Sample calculation of *number of mentions* feature over different domains.

Length features

Length of an ad may provide a good indication of the classification of the ad. A business user is likely to describe his product or service in sufficient detail while ads of non-business users can be very abrupt (as seen in Figure 5.1b). We define two length-based features: length of ad description in characters and in words.

URL based features

We define three URL based features. These are (1) average length of URLs (2) average number of digits in URLs (3) average number of slashes in URLs. The intuition behind these features is that business users are more likely to refer the visitors to their business’ official website for additional details. While it is likely for non-business users to post URLs too (for example, official manufacturer website of a product that the user is trying to sell), such URLs are typically long (since they refer to a location deep down the manufacturer’s website hierarchy) thus containing larger number of slashes (as path separators) and possibly digits. On the other hand, business users are likely to provide a link to the homepage of their official website, which is not expected to be long.

Miscellaneous features

Finally, we also use some boolean (True/False) features. These are (1) is the ad a “wanted” ad (i.e. a product/service is required instead of offered)? (2) is the poster open to trades? (3) is the item or service offered for free? The intuition behind these features is straightforward; businesses are neither likely to post an ad requiring a service (Kijiji has a separate (*jobs*) category for businesses to post hiring

notices) nor expected to be open to barter or giving items for free.

5.5 Dataset Preparation

To prepare dataset for the ads classification task, we were faced with an important question: at which category depth do we classify the ads? Recall from Chapter 3 that an ad category is an n -ary tuple with n indicating the depth of the category. We could perform a business/non-business classification of the ads at a depth of 0 (i.e. same training data for all categories), 1 (i.e. different training data for (*buy and sell*) and (*cars & vehicles*) but same for all their respective sub-categories) and so on. As the depth increases, the task of data collection becomes more time-consuming since the number of categories increases exponentially. Ultimately, we settled on a depth of 1. Our decision was based on the fact that multiple categories may have different terms relevant to different class labels. 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 (*real estate*) and (*pets*) categories respectively (not considered in this work) but might not be very popular with other businesses. However, within a super category (the first element in the category tuple, such as *buy and sell*), the general vocabulary of business and non-business users should not differ very much owing to the categorization in Kijiji.

Therefore, we took a random sample of ads from both (*buy and sell*) and (*cars & vehicles*) categories in the abridged dataset and labeled them manually. As mentioned earlier, distinguishing between the classes for an ad is sometimes confusing, therefore, we labeled an ad into one of the $\{\textit{business}, \textit{non-business}, \textit{unknown}\}$ classes. Table 5.3 shows some statistics of the labeled data.

From Tables 4.1 and 5.3, it is quite clear that while both 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. Finally, the percentage of business ads in (*cars & vehicles*) category

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

Table 5.3: Statistics for dataset prepared for ad classification task.

is more than twice of that for *(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 *(buy and sell)* category.

5.6 Results

5.6.1 Ad Classification

We performed a 10-fold cross-validation experiment to classify ads using the methodology described in Section 5.4. We used both approaches for training classifiers: (1) random under-sampling with an ensemble of classifiers (RUSEC) detailed in Section 4.2.3 and (2) using entire imbalanced data (IMB). The classifiers used are mentioned in Section 4.2.2. We used unigrams and bigrams of ad titles and descriptions for vocabulary and a profile size of 100 for the experiments. Note that while building class profiles, we deliberately chose to ignore the following terms from the vocabulary:

- Numbers; Numbers are special terms not considered relevant to the task of classification. During manual checking, we found only one number that had particular reference to the business class; the toll free number prefix 800 (or 1 800 with country code). Therefore, we kept only this number in the vocabulary.
- Rare terms that occur in two or less ads. Based on the Zipf’s law, only a few terms occur frequently while the majority of terms occur rarely. Remov-

		Business			Non-Business		
		Precision	Recall	F-measure	Precision	Recall	F-measure
<i>(buy and sell)</i>							
RUSEC	LR	0.38	0.87	0.53	0.99	0.81	0.89
	SMO	0.35	0.89	0.5	0.99	0.8	0.89
	RF	0.41	0.89	0.56	0.99	0.82	0.9
IMB	LR	0.84	0.56	0.67	0.96	0.94	0.95
	SMO	0.84	0.45	0.58	0.95	0.95	0.95
	RF	0.81	0.47	0.59	0.95	0.94	0.95
<i>(cars & vehicles)</i>							
RUSEC	LR	0.77	0.91	0.83	0.98	0.9	0.94
	SMO	0.82	0.87	0.84	0.97	0.94	0.96
	RF	0.8	0.89	0.84	0.97	0.92	0.95
IMB	LR	0.96	0.85	0.9	0.97	0.97	0.97
	SMO	0.94	0.79	0.86	0.96	0.97	0.97
	RF	0.97	0.81	0.88	0.96	0.98	0.97

Table 5.4: Results for ad classification. Profile size is set to 100.

ing these terms reduced our vocabulary size extensively and led to a faster processing.

The results are reported in Table 5.4.

First of all, we observe that 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 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 essentially balanced.

However, the reverse 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 previously. 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 balanced data, thus, they tend to over-represent the minority class in their final predictions in comparison to its true underlying distribution.

We notice that with the IMB training, LR (Logistic Regression) performs much better than the other classifiers for both classes in terms of F-measure. For the RUSEC approach, RF (Random Forest) performs best for (*buy and sell*) category and gives nearly good results for (*cars & vehicles*) as SMO which takes the lead here.

Moreover, it can be seen that the recall of the non-business class is much lower for (*buy and sell*) category 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 than (*cars & vehicles*) on the basis of their sub-categories (Section 3.2). Hence, the probability that the top few n-grams of the profile can capture sufficient vocabulary for non-business users in (*buy and sell*) category is lower when compared with (*cars & vehicles*) (recall that for the experiment, profile size is fixed for both categories and is set to 100).

Finally, we notice that the precision of the business class is much lower for (*buy and sell*) than that for (*cars & vehicles*). 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 5.3). Therefore, similar percentage of non-business instances being misclassified will have a much forceful impact on the precision of business users for (*buy and sell*) than (*cars & vehicles*).

5.6.2 Impact of Profile Size

In the previous section, we used a fixed profile size of 100 for all our experiments. A question arises if using a larger number of terms from the profiles can have a positive impact on the classification performance. In this section, we study the effects of varying profile size on ads classification.

We select the best performing classifiers for both training approaches (Random

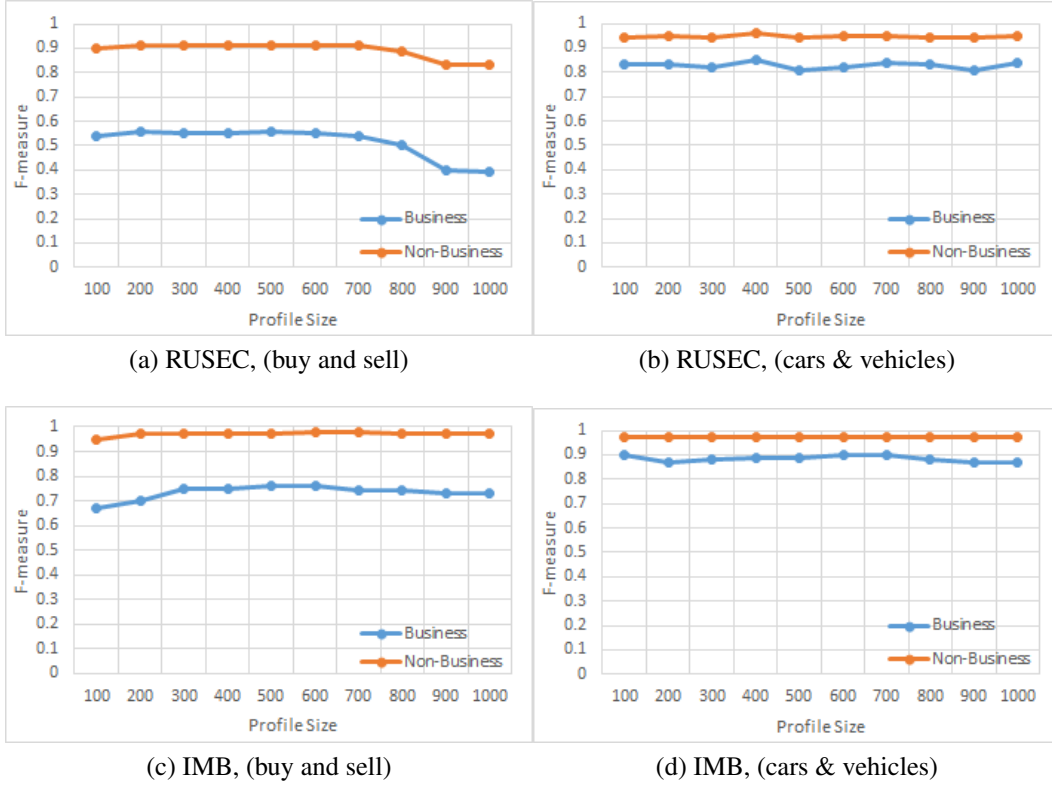


Figure 5.2: Impact of varying profile size on results of ad classification.

Forest for RUSEC and Logistic Regression for IMB) and repeat the same experiment by increasing profile size from 100 to 1,000 using an increment size of 100.

Figure 5.2 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. 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 the part of common users vocabulary rather than being distinctive for a particular class (as they have lower importance scores, Section 5.3).

The only exception to the aforementioned trend is the IMB training for the (*buy and sell*) category. Here, as we increase the profile size, the F-measure increases significantly for the first two iterations. Specifically, F-measure for business class

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

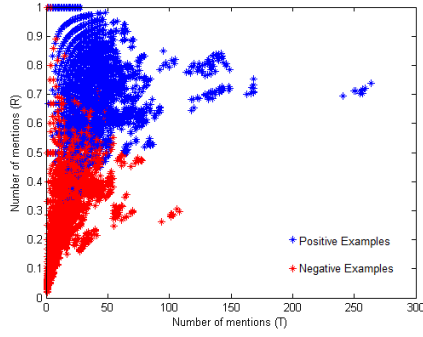
Table 5.5: Feature ranking for ads classification task.

increases by nearly 8%. However, the scores become stable at this stage and experience a slight decline as more terms are incorporated into the profile.

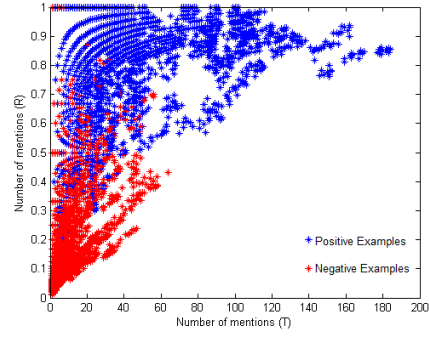
5.6.3 Feature Analysis

In order to find the best performing features, we used Weka to rank all the features using information gain statistic. The profile size was set to 200. We performed a cross-validated ranking, in which the ranks of all features are averaged over the number of folds (10 in our case). Table 5.5 lists the top-10 features for both categories.

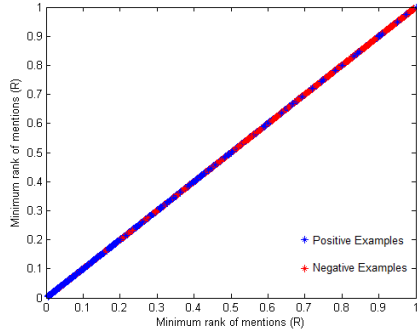
We notice that computing features over domains other than the *T-domain* turned out to be very effective as at least 7 of the top-10 features are from *R*- and Δ -domains for both categories. No statistic from *O-domain* was able to secure a place in the top features, suggesting that the features based on a particular class are more



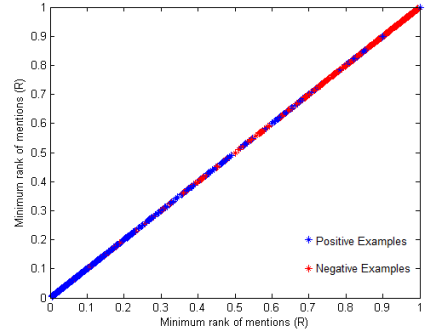
(a) Number of mentions (T) / Number of mentions (R) for (*buy and sell*)



(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*)

Figure 5.3: Distribution of training examples across different feature spaces.

informative. The same is true for *non-profile* features, which contributed only a negligible information gain.

It can be seen from Table 5.5 that *relative number of mentions* is the most impressive feature for both categories, dividing positive and negative examples most accurately. 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 5.3a and 5.3b. While the ads mentioning 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.

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 5.3c and 5.3d). Accordingly, we find that *relative minimum rank of mentions* is an important feature for both categories.

5.7 Classifying users

At this point, let us recall that our original goal was to classify a *user* into *business* or *non-business* classes. In Section 5.1, we observed that the content of users' ads can be very useful in this classification task. Moreover, we noted that due to our definition of business users, the text of only a few ads can provide enough evidence to properly classify a user as business, and for this reason, we classified all the ads of every user into the same classes.

Now that we have devised a technique to give us predictions and associated confidence score for ads classification, a problem to be addressed is how the results of ads classification can be used to predict a class for the user. Needless to mention, the ideal user classification strategy (keeping in view our definition of business users) would be to categorize a user as business if he has posted even a single business ad (as predicted by our classifier). However, we observed that this strategy, while achieving the highest recall, gives a poor F-measure for the business class. This is due to the fact that our ad classification strategy is not perfect (as depicted in Table 5.4) nor we ever expected it to be.

A variant of the user classification strategy described above is to then predict a user as business if more than X% of his ads are inclined towards business class. Figure 5.4 shows the cumulative distribution function of the percentage of business ads against the percentage of users in our dataset for both training approaches: RUSEC and IMB. These results were obtained using the best performing classifiers for each training strategy (i.e., Random Forest for RUSEC and Logistic Regression for IMB) with a fixed profile size of 200 to classify all the ads of every user (except *unknown* ones) in our dataset (Table 4.1). As explained in Section 5.6.1, the IMB approach is much more conservative in predicting an instance as belonging to minority class (*business*) than RUSEC. This is visible from the fact that when

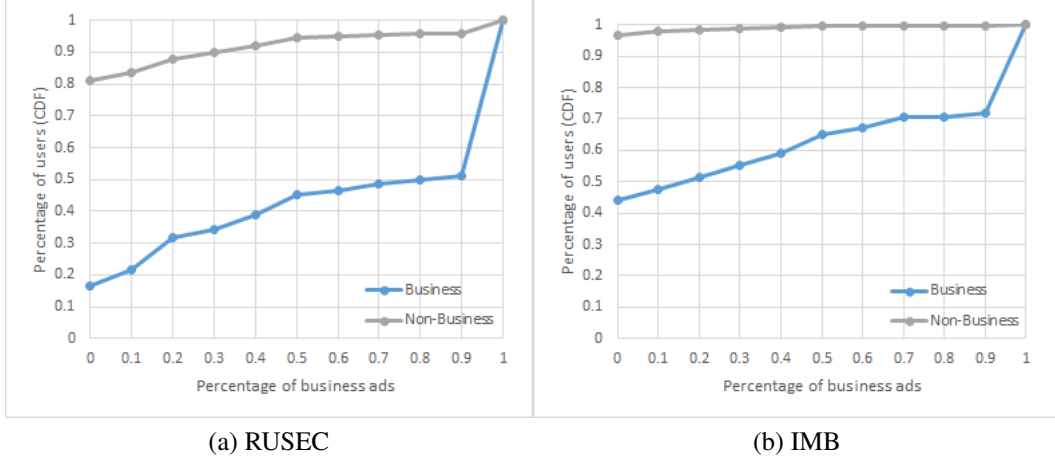


Figure 5.4: CDF of fraction of business ads detected by our classifier against the percentage of users in our dataset.

classified with the IMB strategy, nearly 44% of manually labeled business users are predicted to have no business ad at all, while for RUSEC, the fraction is only 16%. On the other hand, the RUSEC approach predicts that approximately 5% of annotated non-business users have posted more business oriented ads than non-business ones (>0.5), while the number is only 0.6% for the IMB. Keeping these trade-offs in mind and depending on the nature of the application (or analysis), one can classify a user as business only if he has posted more than $X\%$ of ads related to his enterprise. A disadvantage, however, of this user classification strategy is that no meaningful score can be attached to a user indicating how inclined he is towards one class as compared to the other. Due to this reason, we devise an alternative policy to classify users.

To classify ads, we adopted a one-vs-all classification strategy (Section 5.4), under which we trained two classifiers, one for each class. The classifier for a particular class outputs 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. Hence, we associate a score for a label to each user.

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

where $a_{u,c}$ represents the ads of the user u that mention terms from the profile

of class c .

Finally, the user is assigned the class label that gives the highest confidence value.

5.7.1 Experimental Evaluation

We trained the classifiers using both approaches (RUSEC and IMB) with a fixed profile size of 200 and classified ads of all the users in our dataset (Table 4.1). We used the previously mentioned aggregation strategy to predict a class label for each user. To evaluate our method, we used the following baselines for a comparison:

- **Dealer:** As mentioned in Section 3.1, in the Kijiji network, each ad has multiple attributes attached to it. Kijiji allows users to self-identify themselves as businesses through a few such attributes, in return for various benefits. For the categories in our abridged dataset, the relevant attribute is *For Sale By* which can have two possible values *Dealer* or *Owner*. This attribute, however, is not present for all the categories in Kijiji. As far as our dataset is concerned, this attribute is present in the (*cars & vehicles*) and (*buy and sell, furniture*) categories only. Therefore, this baseline will not be able to detect businesses from other sub-categories in (*buy and sell*). Note that for this baseline, we consider the *For Sale By* attribute in “offering” ads only i.e. ads in which an item or service is offered (instead of wanted).
- **Short URL:** We mentioned in Section 5.4 that many users post a URL in their ad description. For business users, this is usually the link to their official website to give viewers more information about their services. On the other hand, for non-business users, it is mostly a link to a page on the manufacturer’s official website detailing the features of the item being sold; as such, those 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/`

		Business			Non-Business		
		Precision	Recall	F-measure	Precision	Recall	F-measure
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	Dealer	0.93	0.17	0.28	0.97	1.00	0.99
	Short URL	0.68	0.16	0.26	0.97	1.00	0.99
	One ad per week	0.04	0.27	0.08	0.97	0.8	0.88
	Weighted Random	0.03	0.03	0.03	0.97	0.97	0.97

Table 5.6: User classification results. Profile size is set to 200.

`~mwaqar/kijiji/` 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:** We expect business users to use Kijiji *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:** 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 4.1), we classify the user as business. Otherwise, the user is classified as non-business.

The result of the user classification task is presented in Table 5.6.

As discussed in Section 5.6.1, 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 clas-

sifiers 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. Note that both 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 or not, we applied paired t-test on results obtained from each classifier and each baseline method. The null hypothesis is: *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.

5.8 Using Unlabeled Data

In the previous section, we used dataset of ads manually labeled (Table 5.3) to classify all the ads of the users and aggregated those results to come up with a class label for each user in our dataset (Table 4.1). Let us consider a scenario where labeled dataset of ads is unavailable and a supervised classification may not be possible. This scenario is not very far-fetched as significant time and effort is required to collect the training data by hand. Moreover, recall from Section 3.2 that unless otherwise stated, for all the experiments in this work, we are considering only two categories (*buy and sell*) and (*cars & vehicles*) as an abridged dataset. If we want to extend our analysis to include other categories, we would have to spend

		Business			Non-Business		
		Precision	Recall	F-measure	Precision	Recall	F-measure
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 5.7: Results for users classification using unlabeled ads dataset. Profile size is set to 200.

significant time to label the data by hand.

To tackle this scenario, we employ a simple bootstrapping heuristic. We provide the system with 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 and vice versa. We thus obtain a labeled dataset to act as the training data for the classification of ads of the users. Therefore, normal ad classification methodology follows.

5.8.1 Experimental Evaluation

We used the classifiers trained using the RUSEC approach with profile size set to 200. We used only 4 n-grams as a seed set: *satisfaction*, *guaranteed*, *priority* and *hours of operation*. Moreover, we also set a limitation 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 limitation was implicit in the manually labeled ads dataset, since while collecting the said dataset, a random sample of all the ads was taken and the probability that extensive ads from a particular user would make into this sample was extremely low owing to the large number of ads in the collection.

The results of user classification using unlabeled ads dataset are mentioned in Table 5.7. Overall, 1,530 and 722 ads were selected from (*buy and sell*) and (*cars & vehicles*) categories respectively. We notice that we achieve a remarkably close F-measures for all the classifiers as compared to when manually labeled ads dataset was utilized (Table 5.6). In fact, Random Forest classifier even exceeds its

F-measure by a single point for both classes.

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.

Chapter 6

Studying Users by Posting Behavior

6.1 Motivation

In the previous chapter, we looked at the problem of user classification into $\{business, non-business\}$ classes by considering each ad individually and primarily by utilizing the text of the posted ads. Specifically, we built a language model for all the classes and computed a confidence score of each ad belonging to a particular class depending on how it mentioned terms from the LM of the respective class. These confidence scores of all the ads by a user were later aggregated to arrive at a label for the user. However, it is not too hard to imagine that in addition to the text posted by the users in their ads, the collective behavior of the user in posting ads itself can convey a lot of information about the user. Consider the ads posted by two users as shown in Figure 6.1. Given the task to classify these users into one of the above-mentioned classes based on this data, one is likely to correctly label user in Figure 6.1a as business while the one in Figure 6.1b as non-business. By re-tracing our thinking which led us to arrive at this decision, we observe that this is primarily because the former is more active on Kijiji (in terms of the number of ads posted), tends to post a large number of distinct items (as opposed to duplicates) and more importantly has a collective theme behind his ads i.e., majority (if not all) of his ads are posted in a category intended to sell cars and trucks. All these factors combined imply that the user is some sort of a car dealer (or involved in a “fix and flip” business) using Kijiji to promote his considerable inventory. On the other hand, the user in Figure 6.1b seldom posts on Kijiji and even then many of his

ads are *reposts* (duplicate ads promoting past listings). Furthermore, the ads lack a collective theme in that they are distributed across various dissimilar categories. Note that while doing this analysis, we did not avail text from titles and descriptions of ads but only to detect identical ads. Thus, it can be seen that besides text, other statistics based on users posting behavior can provide helpful indicators to distinguish between the two user groups. In this chapter, we exploit this dimension and study the behavior of users using their posting patterns.

6.2 Behavioral Features

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 on a classified ad network often provides useful cues as to whether the user is a business. Two frequency-based features w.r.t time are considered: average number of ads per week and standard deviation of number of ads per week. The idea behind the former is to separate users by their activity level on Kijiji while the latter indicates if the their posting activity remains consistent over time or experiences great fluctuations.

Likewise, we also look at inter-arrival time of the ads as an indicator of how actively users utilize the network. The features used here are average inter-arrival time of the ads (in days) and standard deviation of the inter-arrival time of the ads (in days). The motivation is to identify how soon the users return to Kijiji 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 (*empty* categories). The intuition behind these features is to model how closely the ads of the user follow a collective theme and if the user tends to post a large number of ads in a particular category (or set of categories) or if the postings are distributed evenly across various categories.

Visualizing Edmonton Kijiji

[Home](#) | [Browse User Ads](#) | [Browse Ads](#)

[Z75773480](#)

buy and sell
(no ads)

services
(no ads)

cars & vehicles

- GMC SIERRA 4X4 (Steelz on Wheelz 2006 GMC Sierra 1500 4x4 To view all of inventory visit [www.steelzonwheelz.ca. all ...](#))
- 2007 Chevrolet Silverado 1500 4x4 z71 (Steelz on Wheelz 2007 Chevrolet Silverado 1500 4x4 z71 To view all of inventory visit [www.steelzonwh...](#))
- 2003 Ford F-250 (Steelz on Wheelz 2003 Ford F-250 To view all of inventory visit [www.steelzonwheelz.ca. all vehicle c...](#))
- 2002 Chevrolet Blazer 4x4 (Steelz on Wheelz 2002 Chevrolet Blazer 4x4 to view all of inventory visit [www.steelzonwheelz.ca. all...](#))
- 2005 Dodge Ram 1500 (Steelz on Wheelz 2005 Dodge Ram 1500 To view all of inventory visit [www.steelzonwheelz.ca. all vehic...](#))
- 2008 Ford f-150 4x4 (Steelz on Wheelz 2008 Ford f-150 4x4 just spent \$1800 on mechanical. 5.4l xlt, 4x4...comes with 6 mo...)
- 2008 Ford F-150 4x4 (Steelz on Wheelz 2008 Ford F-150 4x4 To view all of inventory visit [www.steelzonwheelz.ca. all vehic...](#))
- 2006 Ford Fusion (Steelz on Wheelz 2006 Ford Fusion Fully loaded, leather, sunroof, a/c, all power options.To view all...)
- 2005 Dodge Heavy duty (Steelz on Wheelz 2005 Dodge Heavy duty To view all of inventory visit [www.steelzonwheelz.ca. all veh...](#))
- 2002 Dodge Dakota (Steelz on Wheelz 2002 Dodge Dakota To view all of inventory visit [www.steelzonwheelz.ca. all vehicle...](#))
- 2004 Ford Ranger (Steelz on Wheelz 2004 Ford Ranger 4x4 nice pickup. call 7807692279 for more info. comes with warrant...)
- 2000 Ford F-150 (Steelz on Wheelz 2000 Ford F-150 To view all of inventory visit [www.steelzonwheelz.ca. all vehicle c...](#))
- 2006 Ford F-150 xlt (Steelz on Wheelz 2006 Ford F-150 xlt To view all of inventory visit [www.steelzonwheelz.ca. all vehic...](#))
- 2002 Chevrolet Silverado 1500 (Steelz on Wheelz 2002 Chevrolet Silverado 1500 To view all of inventory visit [www.steelzonwheelz.ca....](#))
- 2009 Ford F-150 (Steelz on Wheelz 2009 Ford F-150 To view all of inventory visit [www.steelzonwheelz.ca. all vehicle c...](#))

(a) Business (abridged, for complete listing, use system described in Section 3.3 to query User ID Z75773480)

Visualizing Edmonton Kijiji

[Home](#) | [Browse User Ads](#) | [Browse Ads](#)

[Z85504170](#)

buy and sell

- 2" lime stone rocks (Come pick up this rock. I have already got it off the ground. As per the quantity please refer to th...)
- Twin Mattress for FREE (I have a twin mattress to give away for free. It is used. The box spring is old mattress is a few ye...)
- Mattress for free. First come gets it (I have a twin mattress to give away for free. It is used. The box spring is old mattress is a few ye...)

services
(no ads)

cars & vehicles

- Wanted: King Quad 750 AXI Gas Cap (I am looking for a fuel cap for a 2008 king quad 750. The 400 will work plus other years. Please con...)
- Bub BIG BORE pipe (Used Bub Big Bore Pipe. it will fit 02-07 Grizzly 03-06 Kodiak 05-07 King Quad 02-07 Vinson 05-06 50...)
- 2002 Chevrolet Tahoe SUV (So the tahoe has Leather Power seats Air/tilt/cruise Deflectors on the hood and all four doors Towin...)
- 2002 Chevrolet Tahoe SUV Priced to go (So the tahoe has Leather Power seats Air/tilt/cruise Deflectors on the hood and all four doors Towin...)
- BUB BIG BORE SLIP ON PIPE (Used Bub Big Bore Pipe. it will fit 02-07 Grizzly 03-06 Kodiak 05-07 King Quad 02-07 Vinson 05-06 50...)
- 2008 king quad plastics (I have a used full set of plastics that will fit a 2008 king quad. Will also consider trades)
- 2002 Chevrolet Tahoe SUV 4x4 priced to move (So the tahoe has Leather Power seats Air/tilt/cruise Deflectors on the hood and all four doors Towin...)
- 2002 Chevrolet Tahoe SUV (So the tahoe has Leather Power seats Air/tilt/cruise Deflectors on the hood and all four doors Towin...)
- BUB BIG BORE SLIP ON PIPE (Picked this up and found out it does not fit my quad. so selling it at a reduced price.)
- 2008 King Quad 750 AXI plastics (There used and do have a crack as shown in the pic. Will sell for 150 or look at possible trades. It...)
- 2008 Suzuki king quad 750 AXI Plastics (There used and do have a crack as shown in the pic. Will sell for 150 or look at possible trades. It...)

(b) Non-Business

Figure 6.1: Sample Business and Non-Business users.

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), 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 on Kijiji 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 now (or 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

An ad posted on the Kijiji network remains listed for a maximum of 60 days (unless the user chooses to remove it before this period either because the item has been sold or acquired or the user is no longer interested in providing or requiring a service). During the time the ad remains active, the user cannot post a similar ad. Kijiji automatically identifies if a new ad being posted by a user is *similar enough* to one of his ads already listed and prevents from posting a duplicate. After 60 days (or after the ad has been prematurely deleted), the user has an option to list the ad again. It should be noted that by default, the ads returned for a search on Kijiji 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. Even though Kijiji provides various paid

features for the users to highlight or *bump* their ads, 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* of an ad using the concept of w-shingles as introduced by Broder [9]. Specifically, given an ad, we extract unique 1- and 2-shingles from its titles and description. For example, the 2-shingles of the text (*Selling 29 inch TV and selling 29 CDs.*) are (*selling 29*), (*29 inch*), (*inch tv*), (*tv and*), (*and selling*), (*29 cds*). Note that while computing shingles, punctuation marks are eliminated, text is folded to lower-case (for case-independent matching) and only unique shingles are kept. Then, the resemblance r of two ads A and B can be computed using Jaccard similarity coefficient as:

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

where $S(A)$ is the set of 1- and 2-shingles of ad A , 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 manual checking), we conclude that the ad posted later in the pair is a repost of the other one. Note that, a repost cannot have a repost of its own; all such nested reposts are linked to the original ad of the series.

Length features

As length based features, we use the average length of ads description 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.

Wanted ads

Recall from Section 3.1 that ads whose titles begin with the keyword *Wanted* are those in which an item or a service is required by the poster. We use the fraction

	Total	Business	Non-Business	Unknown
Number of users	3,254	110	2,975	169
Percentage of users	-	3.38	91.43	5.19

Table 6.1: Statistics for filtered dataset (users who have posted at least two ads).

of wanted ads as a feature to describe how often a user utilizes the Kijiji network when in need of an item or a service.

6.3 Experiments and Evaluation

6.3.1 Classifying Users Based on Posting Behavior

It does not make sense to define posting patterns for a user who has posted only one ad in the network. Therefore, we filter all the users in the dataset (Table 4.1) to the ones who have listed at least two ads on Kijiji. The statistics for the filtered dataset thus obtained are listed in Table 6.1.

We performed a 10-fold stratified cross-validation on the filtered dataset using both RUSEC (random under-sampling with an ensemble of classifiers, Section 4.2.3) and IMB (imbalanced) training approaches to classify the users into *business* or *non-business* classes. We determined the folds only once and used them throughout the experiments to maintain consistency. We observed in our experiments that when restricted to only top-10 features (based on information gain), we obtain better results as compared to when all the features are used. Hence, for our experiments, we trained the classifier using only the ten most impressive features listed in Table 6.2. The results of our experiment are reported in Table 6.3.

We observe that the results are very disappointing. The IMB training strategy, that yielded most impressive results when using content of ad for user classification, performs poorly here. For two classifiers (SMO and RF), it fails to identify even a single business user correctly whereas for the LR, the results are not very encouraging either. This can be explained, however, based on the nature of imbalance in the data. From Table 6.1, we observe that the users dataset utilized in this experiment is more imbalanced than either of the ads datasets used in the previous experiments (Table 5.3). Since the minority class is only sparingly present, the classification

Feature	Identifier	Average rank
Average length of ad description (in characters)	lic	1.3 ± 0.46
Average length of ad description (in words)	liw	2 ± 0.77
Std. dev. of ads in categories (with <i>empty</i> categories)	stn_apc	2.7 ± 0.46
Std. dev. of ads in categories (without <i>empty</i> categories)	sti_apc	4.3 ± 0.64
Fraction of reposted ads	fra	5.2 ± 0.75
Change in ads per week	ciapw	6.1 ± 1.7
Average length of epoch (in days)	epoch_len	8 ± 1.41
Fraction of unique ads reposted	far	8.3 ± 1.27
Std. dev. of number of ads per week	std_apw	9.4 ± 3.2
Fraction of active time	fat	10.6 ± 2.06

Table 6.2: Ranking of posting behavior features.

		Business			Non-Business		
		Precision	Recall	F-measure	Precision	Recall	F-measure
RUSEC	LR	0.1	0.64	0.18	0.98	0.8	0.88
	SMO	0.09	0.51	0.16	0.98	0.82	0.89
	RF	0.08	0.68	0.14	0.98	0.71	0.82
IMB	LR	0.19	0.04	0.06	0.96	1.0	0.98
	SMO	0.0	0.0	0.0	0.96	1.0	0.98
	RF	0.0	0.0	0.0	0.96	1.0	0.98

Table 6.3: Results for users classification.

algorithm tends to ignore it altogether to achieve a maximum overall accuracy.

Using RUSEC approach to balance the training data helps to predict much more business users correctly, but it is achieved at the cost of a huge decrease in recall of non-business users, which takes its toll on the precision (and consequently the F-measure) of business class. In fact, the results are even less accurate than the ones achieved by *Dealer* and *Short URL* baselines over all users (Table 5.6) in terms of F-measure.

We conclude that the features defined in Section 6.2 to model posting behavior of the users are inadequate to differentiate between business and non-business user groups on Kijiji satisfactorily.

6.3.2 Usage Patterns

We used the aforementioned features in this chapter, defined for modeling the posting behavior of the users, to determine some of the common usage patterns. These

	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%)

Table 6.4: Confusion matrix of the clusters automatically detected by EM algorithm along with the manually annotated data and the percentage of a particular class of users in each cluster. Features used for clustering are listed in Table 6.2.

patterns would not only serve to understand how people most commonly use the network but would also help to comprehend the unexpected results obtained in the previous section. For this purpose, we used the Expectation-Maximization (EM) clustering algorithm (off-the-shelf implementation from Weka [23]) to automatically cluster users using the top-10 posting behavior features (Table 6.2). EM was used since it gives an explicit grouping of the users (as opposed to a dendrogram which has to be manually cut) and can decide how many clusters to create automatically by cross-validation. The number of clusters determined automatically by this method were four. The confusion matrix of the resulting clusters is shown in Table 6.4.

Table 6.5 summarizes the clusters in terms of the features they use. It can be observed that there are significant differences in the posting behavior of users in different clusters. Cluster 1 has the least active users (*inactive* cluster) in terms of all our posting frequency features. The users making up this cluster, on average, remain active for only 1-2 days at a time. In addition, 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 (*active* cluster). While users in the other two groups exhibit values between these two extremes, it can be noted that those in cluster 3 are slightly more active than their counterparts in cluster 2. We refer to the users in these clusters respectively as *active* and *less active*.

Moreover, we observe that ad description lengths and reposting behavior do not always follow the activity trends mentioned previously. Specifically, *less active*

Identifier	Cluster 1 (<i>inactive</i>)	Cluster 2 (<i>less active</i>)	Cluster 3 (<i>active</i>)	Cluster 4 (<i>highly active</i>)
std_apw	0.43 (0.31)	1.26 (0.91)	1.27 (0.79)	2.37 (3.47)
fat	0.04 (0.08)	0.10 (0.12)	0.14 (0.13)	0.25 (0.24)
fra	0.06 (0.14)	0.01 (0.14)	0.31 (0.16)	0.57 (0.25)
far	0.10 (0.25)	0.01 (0.14)	0.32 (0.20)	0.63 (0.35)
lic	272.34 (190.48)	188.45 (83.66)	269.87 (162.80)	633.11 (612.65)
liw	48.44 (33.07)	33.93 (15.22)	48.18 (29.00)	107.90 (99.04)
stn_apc	0.14 (0.05)	0.31 (0.17)	0.48 (0.28)	1.65 (6.27)
sti_apc	0.15 (0.25)	0.88 (0.84)	1.60 (1.33)	6.45 (35.51)
epoch_len	1.49 (1.06)	3.48 (2.66)	4.61 (3.30)	10.37 (19.04)
ciapw	9.20 (5.82)	5.11 (2.97)	3.91 (2.13)	3.90 (3.91)

Table 6.5: Mean value of features for clusters (Table 6.4). Standard deviations are shown inside brackets. Feature identifiers are listed in Table 6.2.

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 Problem By Usage Patterns

Using the top-ranked features from those in Section 6.2, we enumerated four distinct usage patterns of Kijiji. From Table 6.4, we notice that both business and non-business users exhibit all the patterns. However, 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, are uti-

lizing Kijiji to announce special promotions and discounts. Other types of business users found in this group are also characterized by low activity, for instance small businesses who should be promoting their offerings in Kijiji's (*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 using Kijiji to give a summary of their services and to promote their official web presence, businesses with time-bound advertisements like a summer camp inviting registration applications etc.

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 viewership of their ads via re-

posting. 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.

From the above discussion, we conclude that although certain user groups have more affinity to a particular usage pattern, contrary to what we expected, it is not exclusive to that group only since a significant fraction of members of the other group 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 as reported in Section 6.3.1; 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 gain a better understanding of how they utilize a classified ad network.

Chapter 7

Analysis of the Classified Ad Network

In the preceding chapters, we modeled the classified ad network from the perspective of *business* and *non-business* users with a focus on distinguishing the users belonging to both classes. Particularly, in Chapter 5, we studied this problem by utilizing the content of the ads that the users post. On the other hand, in Chapter 6, we sought to achieve the distinction between the two user groups on the basis of the collective posting behavior of the users. In this chapter, however, we shift our focus, and analyze the Kijiji classified ad network from various dimensions. Moreover, wherever meaningful, we will use the previously obtained results and examine the results for both user groups separately.

Experimental Setup

For all the experiments in this chapter, we will utilize the entire users dataset as opposed to randomly sampling a few users, a technique we adopted for the experiments in the previous chapters. Moreover, whenever we need to determine the labels for the users, we will classify them using the strategy based on the content of their ads (Chapter 5) using the initial training dataset of Table 5.3. For this purpose, we will utilize the imbalanced (IMB) training approach with Random Forest (RF) classifier and a fixed profile size of 200, the combination that yielded the best results in terms of F-measure (Table 5.6).

7.1 Temporal Changes in User Profiles

Let us begin by recalling that each ad of a user on the Kijiji network is posted in a specific category and a user can have multiple ads in different categories all active (listed) simultaneously. In this section, we investigate how amenable users are to exhibit similar posting trend in terms of number of ads in different categories over time. Specifically, we probe if users tend to post consistently in a particular set of categories or if the concentration of their ads as reflected through their numbers in different categories varies significantly with the passage of time. We study this behavior in the context of the two user groups: business and non-business.

For this purpose, we need to represent the popularity (or weights) of different categories for a particular user during a specific time interval. Similar to Abel et al. [2], we adopt a vector-space model for this purpose and construct *user profiles*.

Definition 1 (User Profile): *The profile of a user $u \in U$ is a set of weighted categories where the weight of a category $c \in C$ for u is computed by a certain function w .*

$$P(u) = \{(c, w(u, c)) \mid c \in C, u \in U\}$$

where C and U denote the set of categories and users respectively.

We use a frequency based scheme to compute the weights of the user profiles. Specifically, for a user u , the weight of a category c is determined by the number of ads u has posted in c . We then normalize the profiles so that the sum of weights in a profile is equal to 1 i.e., $\sum_{c \in C} w(u, c) = 1$.

With this definition, our task is transformed to determining how do user profiles evolve over time. For this reason, we determine the profiles of users for different intervals and use the Manhattan (or L_1) distance to compute the difference between profiles in vector representation (similar to [2]). Using the Manhattan distance, the difference in the profiles of a user u over different intervals x and y is computed as $d(P_x(u), P_y(u)) = \sum_i |P_{x,i}(u) - P_{y,i}(u)|$. Since the profiles are normalized, this distance ranges in $[0..2]$. The higher the distance, the more different are the two

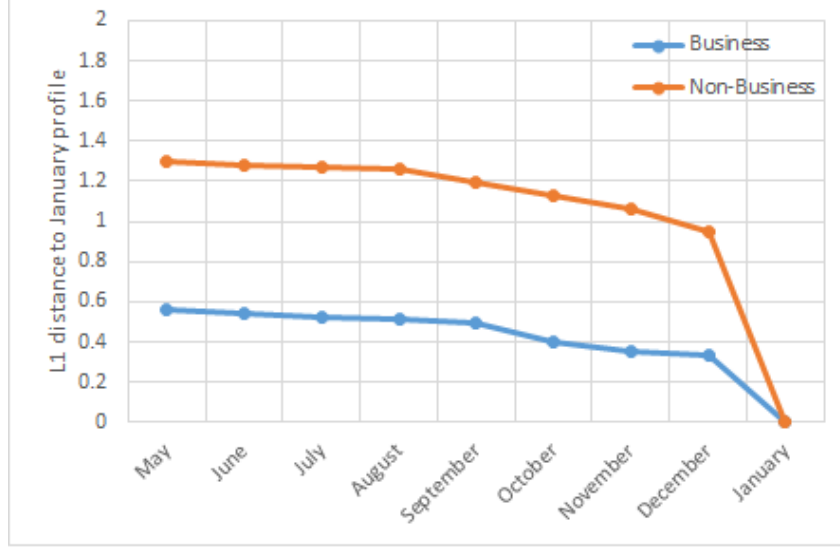


Figure 7.1: Temporal Changes in User Profiles.

profiles $P_x(u)$ and $P_y(u)$ and if both profiles are exactly the same, then the distance is zero.

We use the above-mentioned procedure to study the changes in profiles of both business and non-business users. The labels for the users were determined using the procedure described in the previous section. For the experiment, we used only those users who posted at least one ad in every month covered by our dataset. Recall from Section 3.2 that our dataset spans over 9 months, covering the ads posted between May 2013 and January 2014.

Figure 7.1 shows the average Manhattan distance between user profiles constructed for January 2014 and the profiles of the same users created based on their activity in the network during a certain month in the past. The figure shows that for all the months, the L_1 distance for non-business users is more than twice the same for business users. In other words, non-business users exhibit continuously stronger changes in their profiles than businesses. This result is not unexpected, rather it conforms to our intuitive thinking. Since private individuals tend to use the classified ad network only when a particular need arises, 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 boosting 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 the user groups, the L_1 distance gradually decreases over time. For non-business users, this is expected based on the reasoning mentioned above. However, a similar behavior (albeit smaller in scale) in the profiles of business users may be due to one or any combination of these factors: (1) the presence of non-business ads since our definition of businesses allows them to utilize the network for personal use as well (Section 4.1) (2) incorrect classification of users since, for the experimental setup we are using, the precision of business class is 0.73 (Table 5.6), and (3) seasonal trends and change in the inventory/services.

7.2 Distinctive Categories for User Groups

Given a way to label a user as *business* or *non-business*, in this section, 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 profiles of both classes (Section 5.3). Recall that when constructing the class profiles, we used the KL-divergence and computed the importance score of a term t for class label l using Equation 5.1. Appropriately, the terms having higher importance scores deviated the most from the common vocabulary and became more important (or relevant) to a particular class. Using a similar technique, we determine the category model (CM) for both classes and compute the importance score of each ad category c for a particular class label l using the following equation. The categories having larger such values are then the most distinctive for the respective class.

$$I(c, l) = P(c | l) \ln \frac{P(c | l)}{P(c | D)} \quad (7.1)$$

where D represents the entire corpus i.e., the collection of all users.

The marginal probabilities of categories for the entire corpus $P(c | D)$ and class corpora (collection of business/non-business users) $P(c | l)$ is calculated using the similar *tf-idf* weighting scheme as detailed in Section 5.3, the difference being that $\overline{tf}(c, D)$ now represents the average number of ads posted in category c by the users in D and $df(c, D)$ denotes the number of users in D who posted an ad in category c .

S. No.	Categories
<i>Business</i>	
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)
<i>Non-Business</i>	
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)

Table 7.1: Distinctive categories for user groups.

Table 7.1 lists the top-20 distinctive categories thus obtained for the two classes. We observe that there are significant and meaningful differences between both user groups. Particularly, for businesses, we notice the presence of a large number of *service-oriented* categories (computer services, automotive services, home renovation services and cell phone services) and other *business-oriented* categories (business/industrial, heavy equipment and cargo trailers).

Moreover, it can be observed that the majority of most 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 trend is not surprising, since while many enterprises related to (*buy and sell*) use Kijiji for advocating their offerings, the fraction of such business users and the percentage of their postings are dwarfed by the overwhelming number of non-businesses in most of these categories. Due to this fact, the only (*buy and sell*) 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 (couches/futons and chairs/recliners) are ranked as the top distinctive ones for non-business class.

7.3 Temporal Changes in Popular Categories for User Groups

In the previous section, we sought to determine which categories are the most “distinctive” for a particular user group. However, this does not tell us which categories are especially popular with the respective users during a particular time interval as compared to the entire duration. This is the object of our analysis in this section.

Not surprisingly, the strategy we adopt for this task is a simple variant of the one we used in the previous section. For users belonging to a specific label l , we determine the marginal probability of an ad category c over the entire duration of

the dataset $P(c \mid l)$ as well as for a particular time interval x as $P_x(c \mid l)$. These probabilities are computed using the identical *tf-idf* scheme as described previously. The importance score of category c at interval x , computed using the following formula, then tells us how much the postings in c during interval x differ from the overall trend for users belonging to the same class l .

$$I_x(c, l) = P_x(c \mid l) \ln \frac{P_x(c \mid l)}{P(c \mid l)} \quad (7.2)$$

We use the above strategy to determine the 10 most popular categories for each month from May 2013 to January 2014 as compared to the entire time span of the dataset for both the user groups. Results for non-business users are reported in Table 7.2. Studying these results reveals certain interesting trends. Specifically, 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, thereby, indicating that a large number of university students used the Kijiji network to buy and sell used books during these periods. Moreover, we observe that many furniture categories experienced a lot of activity than usual during the same period (August, September). This can also be explained by the fact that this period sees a large influx of newly admitted international students and an exodus of newly graduated students into and out of Edmonton who use the Kijiji network to secure or dispose off the necessary furniture for or from their residences respectively.

Likewise, costumes category sees a higher than usual traction during the month of October, which concurs with the Halloween season, when many are shopping for Halloween costumes. This is reinforced by the fact that many other clothing categories became popular with the non-business users around the same period.

In the same way, we spot many other seasonal trends. For example, categories for garage sales, lawnmowers and leaf blowers experience great activity at the onset of the summer season in Edmonton; electronics, jewellery and watches become especially popular during Christmas (December) and start to lose traction afterward; categories dedicated to video game consoles like Playstation and XBox face higher than usual posted ads around November 2013 when newer versions of both

Categories	M	J	J	A	S	O	N	D	J
(bs, art/collectibles)	-	7	-	9	-	8	10	6	9
(bs, baby items, strollers/carriers/car seats)	-	-	-	-	8	-	-	-	-
(bs, books)	-	-	-	2	1	-	-	-	1
(bs, clothing, costumes)	-	-	-	-	-	3	-	-	-
(bs, clothing, kids/youth)	-	-	-	-	5	6	-	-	-
(bs, clothing, men's)	-	-	-	-	-	9	-	-	-
(bs, clothing, women's - tops/outerwear)	-	-	-	-	-	5	3	8	7
(bs, computer accessories)	9	5	3	6	3	-	-	-	-
(bs, computers)	5	4	2	4	4	-	-	-	-
(bs, computers, laptops)	-	-	-	-	-	-	-	9	5
(bs, electronics)	-	-	-	-	-	-	-	4	6
(bs, furniture, beds/mattresses)	-	-	-	8	-	-	-	-	-
(bs, furniture, chairs/recliners)	-	-	-	-	10	-	-	-	-
(bs, furniture, couches/futons)	-	-	-	10	-	-	-	-	-
(bs, garage sales)	-	3	4	3	9	-	-	-	-
(bs, home - indoor, home decor/accents)	-	-	-	-	-	7	9	-	-
(bs, home - outdoor, lawnmowers/leaf blowers)	7	-	-	-	-	-	-	-	-
(bs, jewellery/watches)	-	-	-	-	-	-	-	3	8
(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	6	8	-	-	-	-	-	-
(bs, sporting goods/exercise, hockey)	-	-	-	-	7	-	-	-	-
(bs, sporting goods/exercise, snowboard)	-	-	-	-	-	-	5	-	-
(bs, tickets)	-	-	-	-	-	4	4	2	4
(bs, toys/games)	-	9	6	5	6	-	-	7	-
(bs, video games/consoles, Sony Playstation 3)	-	-	-	-	-	-	8	-	-
(bs, video games/consoles, Sony Playstation 4)	-	-	-	-	-	-	-	10	-
(bs, video games/consoles, XBOX 360)	-	-	-	-	-	-	7	-	-
(cv, ATVs/snowmobiles, ATVs)	-	-	10	-	-	-	-	-	-
(cv, ATVs/snowmobiles, snowmobiles)	-	-	-	-	-	10	6	5	3
(cv, ATVs/snowmobiles, snowmobiles parts/trailers/accessories)	-	-	-	-	-	-	-	-	10
(cv, RVs/campers/trailers, RVs/motorhomes)	8	-	-	-	-	-	-	-	-
(cv, RVs/campers/trailers, travel trailers/campers)	3	2	5	7	-	-	-	-	-
(cv, auto parts/tires, other parts/accessories)	-	10	-	-	-	-	-	-	-
(cv, auto parts/tires, tires/rims)	-	-	-	-	-	2	2	-	-
(cv, boats/watercraft, powerboats/motorboats)	6	8	7	-	-	-	-	-	-
(cv, classic cars)	-	-	9	-	-	-	-	-	-
(cv, motorcycles, street/cruisers/choppers)	10	-	-	-	-	-	-	-	-

Table 7.2: Popular categories for non-business users over time. Column headers M-J represent months from May 2013 to January 2014. “bs” and “cv” denote *buy and sell* and *cars & vehicles* respectively.

Categories	M	J	J	A	S	O	N	D	J
(bs, art/collectibles)	-	5	2	2	2	3	2	2	2
(bs, business/industrial)	8	-	9	-	6	-	-	-	-
(bs, computer accessories)	-	6	10	10	8	-	-	-	-
(bs, computer accessories, services (training/repair))	-	-	-	-	-	8	-	-	-
(bs, computers)	4	2	4	3	5	-	-	-	-
(bs, computers, desktop computers)	-	-	-	-	-	-	9	-	5
(bs, computers, laptops)	-	-	-	-	-	-	4	4	4
(bs, electronics)	10	-	-	-	-	-	-	-	-
(bs, furniture, beds/mattresses)	-	-	6	4	-	-	5	3	8
(bs, furniture, couches/futons)	-	-	-	-	-	-	-	7	10
(bs, home - indoor, home decor/accents)	-	-	-	-	9	6	-	-	-
(bs, home - outdoor, lawnmowers/leaf blowers)	-	10	-	9	-	-	-	-	-
(bs, home renovation materials, floors/walls)	-	-	-	-	-	-	8	9	7
(bs, jewellery/watches)	-	-	-	-	-	-	-	6	-
(bs, other)	6	-	-	-	-	-	-	-	-
(bs, phones, cell phones)	-	-	-	-	10	4	7	10	9
(bs, phones/PDAs/iPods)	1	-	-	-	-	-	-	-	-
(bs, phones/tablets)	2	1	1	1	7	-	-	-	-
(bs, tickets)	-	-	-	-	1	1	1	1	1
(cv, ATVs/snowmobiles, ATVs parts/trailers/accessories)	-	3	8	5	4	7	-	-	-
(cv, RVs/campers/trailers, RVs/motorhomes)	-	-	3	8	-	-	-	-	-
(cv, RVs/campers/trailers, cargo/utility trailers)	-	4	7	6	3	5	6	8	6
(cv, RVs/campers/trailers, parts/accessories)	9	7	-	-	-	-	-	-	-
(cv, RVs/campers/trailers, travel trailers/campers)	3	-	5	7	-	-	-	-	-
(cv, auto parts/tires, other parts/accessories)	-	-	-	-	-	9	10	5	3
(cv, auto parts/tires, tires/rims)	7	8	-	-	-	2	3	-	-
(cv, boats/watercraft, powerboats/motorboats)	5	-	-	-	-	-	-	-	-
(cv, motorcycles, motorcycle parts/accessories)	-	9	-	-	-	10	-	-	-

Table 7.3: Popular categories for business users over time. Column headers M-J represent months from May 2013 to January 2014. “bs” and “cv” denote *buy and sell* and *cars & vehicles* respectively.

the consoles were released; 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 and exercise categories also reveal interesting seasonal trends. We observe that the category pertaining to golf encounters a surge during summer, the one related to snowboards at the arrival of winter while the same for (ice) hockey in September, which coincides with the beginning of the hockey season for many parents; a similar trend is observed in tickets category during the same period most probably due to interest in tickets for Edmonton Oilers games. The same can also be spotted in many (*cars & vehicles*) categories, as we note that travel trailers, campers, boats, RVs, ATVs and motorcycles 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 (Table 7.3). 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 summer and winter seasons offering their services; activity in RVs, motorhomes, travel trailers, campers, boats, lawnmowers and leaf blowers categories sees an upward jump during summer etc.

At the same time, we can also observe the effect produced by discontinued or newly introduced categories in Kijiji over time. For example, it can be seen that (*buy and sell, phones/tablets*) category was initially very popular with both classes of users. However, around September, it disappears from the ranking and never emerges later. The reason for this behavior is that it was discontinued by Kijiji 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. The same trend can also be witnessed for some other categories

in the results.

7.4 Distinctive Categories for Locations

In Section 7.2, we enumerated categories that are distinctive for a particular user group as compared to the other. In this section, we seek to observe the same phenomenon, but for different localities. In other words, our goal is to find the categories that are the most distinctive or unusual for a particular location.

The strategy that we adopt here is identical to the one used by Backstrom et al. [4]. We model the ads being posted in various categories as Bernoulli trials; the trial is a success if an ad is posted in a particular category c and 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 arrival of ads in categories is independent of each other, the probability of s_x (number of ads posted in category c from location x) successes is given by:

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

The categories having the lowest values of this probability are the ones who differ most significantly at x from their global background rate and thus are the most distinctive categories for location x .

For the experiment, we considered the entire dataset instead of the abridged dataset (Section 3.2) i.e., all the categories in the Kijiji network. For different locations, we utilize the postal code map of Edmonton city issued by Canada Post and shown in Figure 7.2. From the figure, we observe that the entire city of Edmonton has been divided into 38 neighborhoods, each of which starts with a specific 3-digit postal code. To obtain the appropriate neighborhood for each ad, we utilized the *Address* attribute and extracted the postal code present in this field using regular expressions. The ads which did not contain this attribute or had no postal code mentioned in the address were ignored. We found that this was the case for

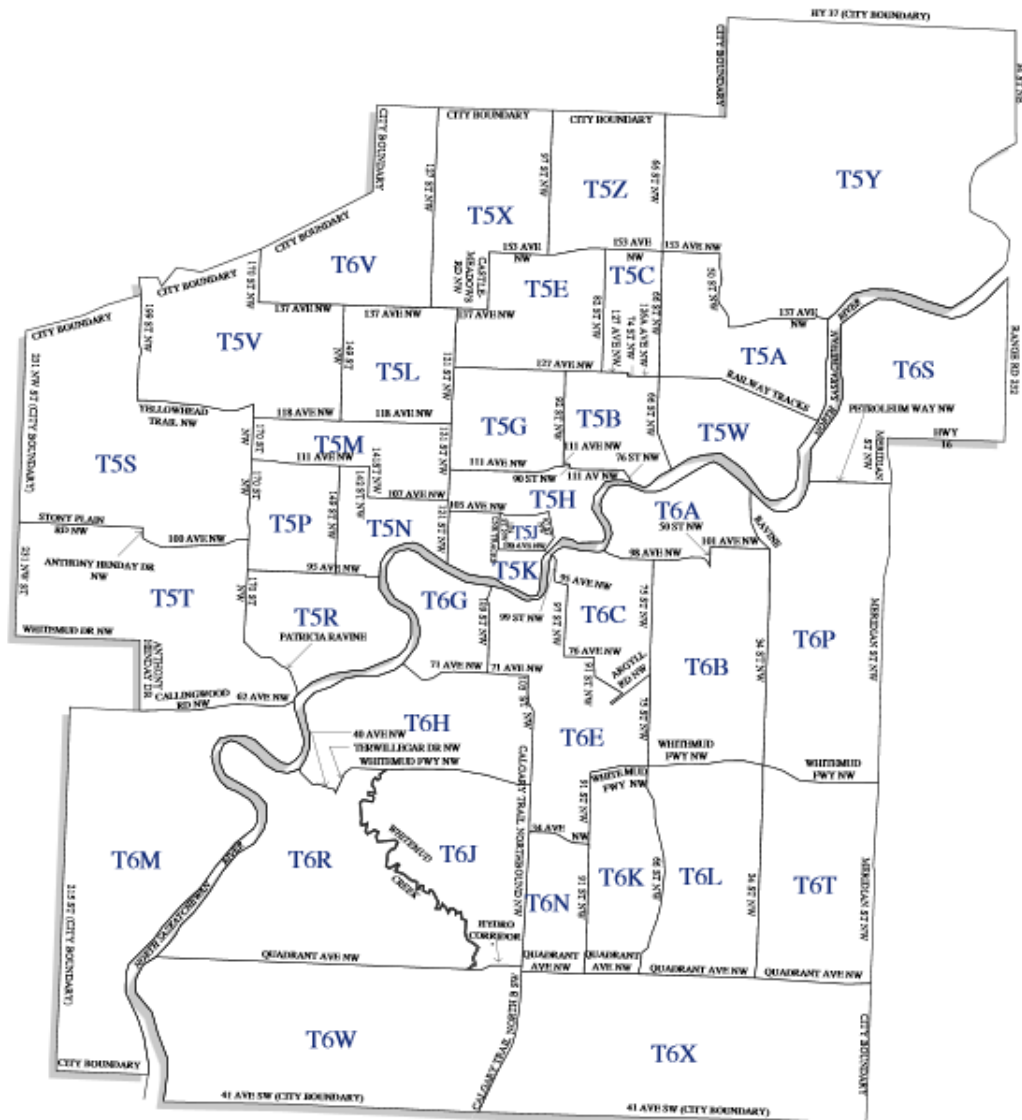


Figure 7.2: Edmonton's Postal Code Map. © Canada Post, 2001.

nearly 24% of the ads, hence, an overwhelming majority of the ads was used in the computation.

The top-7 results for some of the neighborhoods are reported in Table 7.4 (note that the table is divided into two parts on separate pages). 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 in which we would expect great activity by the students (for example phones, books, room rental and clothing). This is not surprising since University of Alberta is situated in this postal code and accordingly the area is inhabited by many local and international students. Since most of the students tend to live closer to the campus, we observe that books category is also popular in the nearby T5K, T5J and T5H communities. This is also reinforced by 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 which is 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, T6A and T6C areas attract a lot of postings from apartments/condos, room rental and house rental categories. Similarly, we observe tickets category listed as distinctive for T5B neighborhood which houses Rexall Place, the home of Edmonton Oilers hockey team.

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 (T6P, T6S, T6N, T5S, T5V), we notice that the jobs become more physically demanding (construction and general labor).

S. No.	Categories
1.	(cv, cars & trucks)
2.	(bs, furniture, couches/futons)
3.	(bs, tickets)
4.	(bs, cds/dvds/blu-ray)
5.	(cv, auto parts/tires, tires/rims)
6.	(bs, other)
7.	(bs, furniture, beds/mattresses)

(a) T5B

S. No.	Categories
1.	(cv, cars & trucks)
2.	(re, apartments/condos, 1 bed-room)
3.	(bs, electronics)
4.	(re, apartments/condos, 2 bed-room)
5.	(bs, tickets)
6.	(re, room rental, roommates)
7.	(bs, books)

(b) T5H

S. No.	Categories
1.	(jobs, sales/retail sales)
2.	(jobs, office mgr/receptionist)
3.	(jobs, bar/food/hospitality)
4.	(jobs, customer service)
5.	(bs, tickets)
6.	(jobs, general labour)
7.	(bs, books)

(c) T5J

S. No.	Categories
1.	(bs, tickets)
2.	(cv, cars & trucks)
3.	(bs, books)
4.	(re, apartments/condos, 1 bed-room)
5.	(bs, electronics)
6.	(bs, sporting goods/exercise)
7.	(re, condos for sale)

(d) T5K

S. No.	Categories
1.	(cv, cars & trucks)
2.	(jobs, construction/trades)
3.	(cv, auto parts/tires, tires/rims)
4.	(jobs, general labour)
5.	(jobs, driver/security)
6.	(bs, art/collectibles)
7.	(bs, furniture, beds/mattresses)

(e) T5S

S. No.	Categories
1.	(jobs, construction/trades)
2.	(cv, cars & trucks)
3.	(bs, furniture, beds/mattresses)
4.	(jobs, general labour)
5.	(bs, furniture, dining tables and sets)
6.	(jobs, driver/security)
7.	(bs, furniture, chairs/recliners)

(f) T5V

Table 7.4: (1/2) Distinctive categories for various neighborhoods. “bs”, “cv” and “re” stand for *buy and sell*, *cars & vehicles* and *real estate* respectively.

S. No.	Categories	S. No.	Categories
1.	(bs, tickets)	1.	(cv, cars & trucks)
2.	(cv, cars & trucks)	2.	(re, house rental)
3.	(bs, art/collectibles)	3.	(bs, tickets)
4.	(bs, sporting goods/exercise)	4.	(bs, sporting goods/exercise)
5.	(bs, video games/consoles)	5.	(bs, electronics)
6.	(bs, electronics)	6.	(bs, art/collectibles)
7.	(re, house rental)	7.	(re, room rental/roommates)

(g) T6A

(h) T6C

S. No.	Categories	S. No.	Categories
1.	(bs, phones, cell phones)	1.	(jobs, construction/trades)
2.	(bs, books)	2.	(cv, cars & trucks)
3.	(bs, phones/tablets)	3.	(cv, motorcycles, motorcycle parts/accessories)
4.	(re, room rental/roommates)	4.	(bs, other)
5.	(cv, cars & trucks)	5.	(bs, business/industrial)
6.	(bs, clothing, women's - tops/outerwear)	6.	(jobs, sales/retail sales)
7.	(bs, tickets)	7.	(services, health/beauty)

(i) T6G

(j) T6N

S. No.	Categories	S. No.	Categories
1.	(jobs, construction/trades)	1.	(jobs, construction/trades)
2.	(cv, auto parts/tires, auto body parts)	2.	(jobs, general labour)
3.	(re, houses for sale)	3.	(cv, auto parts/tires, tires/rims)
4.	(jobs, general labour)	4.	(jobs, driver/security)
5.	(jobs, driver/security)	5.	(cv, cars & trucks)
6.	(cv, cars & trucks)	6.	(bs, home - outdoor)
7.	(bs, art/collectibles)	7.	(bs, business/industrial)

(k) T6P

(l) T6S

Table 7.4: (2/2) Distinctive categories for various neighborhoods. “bs”, “cv” and “re” stand for *buy and sell*, *cars & vehicles* and *real estate* respectively.

Chapter 8

Conclusions

In this research, we seek to characterize users in a classified ad network and to use the results thus obtained to further analyze the network. Specifically, we considered the problem of classification of Kijiji users into either *business* or *non-business* classes. We showed that the content of the ads posted by the users can provide significant clues regarding the label of the user. Moreover, since it is possible for the business users to avail the network for their personal use too (as per our definition of business users, Section 4.1), we noticed that it may not be necessary to examine all the ads of a user, and only a small subset of users ads may contain enough evidence on its own to label the user as a business. Accordingly, we determined the affinity of each ad to both classes based on their language models and aggregated the affinity scores to predict a class label for the user. Our experiments showed this to be an effective strategy for classifying users, achieving significant improvements over various baselines. We also studied the impact of the profile size on the classification performance and noted that a set of highly ranked terms in the profile are most informative and incorporating more terms usually leads to a decline in the results. In addition, in the absence of labeled data for training, we revealed that a simple semi-supervised setting with only a few n-grams as the seed set can be an effective strategy for user classification giving nearly similar results in terms of F-measure.

We also studied the same problem from another perspective: the posting behavior of the users. Using the collective behavioral features of a user in posting ads, we enumerated four distinct usage patterns of Kijiji users and studied the distribution of business and non-business users in these clusters. Our examination revealed that

a sizeable number of users from both user groups validly manifest all the patterns, due to which the aforementioned features are inadequate for the classification task.

Finally, we used the results of user classification to analyze the network from various dimensions. We showed that with the passage of time, the postings of non-business users in different categories vary significantly as compared to the business users who tend to post more consistently in a specific set of categories. Moreover, we determined the distinctive categories for both user groups and observed, as expected, the presence of a large number of service- and other business-oriented categories for businesses. We also showed that the popularity of different categories for both user groups manifests various seasonal trends. Finally, we noted various interesting results when enumerating distinctive categories by Edmonton city's neighborhoods.

8.1 Future Work

In general, distinguishing between *business* and *non-business* users on Kijiji is a challenging task. While we were able to leverage the content of the ads to obtain an effective distinction between the user groups, we believe the results can be further improved by using various features that were not part of our crawled dataset. Specifically, features based on HTML text in ads descriptions, number of views for an ad, attached pictures and usage of paid Kijiji perks (like bumping or highlighting an ad etc.) can be investigated for this purpose.

For nearly all the experiments in this work, we used the abridged dataset, containing only (*buy and sell*) and (*cars & vehicles*) categories. A direction for future research can be to utilize the data from other categories for user classification and subsequent analysis of the network, some of which (for example (*jobs*) and (*services*)) may not have the same degree of imbalanced data as we witnessed for the categories in abridged dataset. Likewise, applying our methodology and validating our results on data obtained from other classified ad networks or different regions in Kijiji could also be an interesting exercise.

An important application of this research could be to integrate the results of user

classification periodically into the Kijiji website via a browser extension (addon or plugin) so that the users can be informed of the nature of the user in real time. More generally, a recommendation system taking into account user label and possibly many other dimensions (for example, history of the user in selling items from a particular category) can be made available to the users.

As a further application of this research, many more aspects of the classified ad network can be analyzed, possibly by considering users as belonging to *business* and *non-business* classes. Predicting how users determine 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.

Bibliography

- [1] K. Aas and L. Eikvil. Text categorisation: A survey. *Raport NR*, 941, 1999.
- [2] 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.
- [3] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.
- [4] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *Proceedings of the 17th international conference on World Wide Web*, pages 357–366. ACM, 2008.
- [5] S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman. Temporal analysis of a very large topically categorized web query log. *Journal of the American Society for Information Science and Technology*, 58(2):166–178, 2007.
- [6] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12, 2010.
- [7] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 620–627. ACM, 2009.
- [8] M. W. Berry and M. Castellanos. Survey of text mining. *Computing Reviews*, 45(9):548, 2004.
- [9] A. Z. Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings*, pages 21–29. IEEE, 1997.
- [10] J. Chan and A. Ghose. Internets dirty secret: assessing the impact of online intermediaries on hiv transmission. *Chan J., and Ghose A.,” Internets Dirty Secret: Assessing the Impact of Online Intermediaries on HIV Transmission”, MIS Quarterly (Forthcoming)*, 2013.
- [11] C. Chen, K. Wu, V. Srinivasan, R. Bharadwaj, et al. The best answers? think twice: Online detection of commercial campaigns in the cqa forums. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 458–465. IEEE, 2013.

- [12] C. Chen, K. Wu, V. Srinivasan, and X. Zhang. Battling the internet water army: Detection of hidden paid posters. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 116–120. ACM, 2013.
- [13] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.
- [14] R. Cooley. Classification of news stories using support vector machines. In *Proc. 16th International Joint Conference on Artificial Intelligence Text Mining Workshop*. Citeseer, 1999.
- [15] H. K. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li. Detecting online commercial intention (oci). In *Proceedings of the 15th international conference on World Wide Web*, pages 829–837. ACM, 2006.
- [16] C. Eickhoff, P. Serdyukov, and A. P. de Vries. Web page classification on child suitability. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1425–1428. ACM, 2010.
- [17] A. Estabrooks, T. Jo, and N. Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36, 2004.
- [18] D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine learning*, 2(2):139–172, 1987.
- [19] J. A. Fries, A. M. Segre, and P. M. Polgreen. Using online classified ads to identify the geographic footprints of anonymous, casual sex-seeking individuals. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 402–410. IEEE, 2012.
- [20] Q. Gao, F. Abel, G.-J. Houben, and Y. Yu. A comparative study of users microblogging behavior on sina weibo and twitter. In *User modeling, adaptation, and personalization*, pages 88–101. Springer, 2012.
- [21] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 369–378. ACM, 2009.
- [22] O. L. Haimson, J. R. Brubaker, and G. R. Hayes. Ddfseeks same: sexual health-related language in online personal ads for men who have sex with men. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1615–1624. ACM, 2014.
- [23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.
- [24] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management*, 44(3):1251–1266, 2008.

- [25] B. J. Jansen, A. Spink, and J. Pedersen. A temporal comparison of altavista web searching. *Journal of the American Society for Information Science and Technology*, 56(6):559–570, 2005.
- [26] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information processing & management*, 36(2):207–227, 2000.
- [27] B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell system technical journal*, 49(2):291–307, 1970.
- [28] W. Klement, S. Wilk, W. Michaowski, and S. Matwin. Dealing with severely imbalanced data. *ICEC09*, page 14, 2009.
- [29] K. Kroft and D. G. Pope. Does online search crowd out traditional search and improve matching efficiency? evidence from craigslist. *Journal of Labor Economics*, 32(2):259–303, 2014.
- [30] J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 31–40. ACM, 2010.
- [31] A. Makazhanov, D. Rafiei, and M. Waqar. Predicting political preference of twitter users. *Social Network Analysis and Mining*, 4(1):1–15, 2014.
- [32] G. Mishne. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, volume 19, 2005.
- [33] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
- [34] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.
- [35] J. Orwant. Heterogeneous learning in the doppelgänger user modeling system. *User Modeling and User-Adapted Interaction*, 4(2):107–130, 1994.
- [36] A. OSullivan and S. M. Sheffrin. *Economics: Principles in action*. Boston, Mass.: Pearson/Prentice Hall, 2007.
- [37] G. Packard, S. G. Moore, and B. McFerran. How can ”I” help ”You”? The impact of personal pronoun use in customer-firm agent interactions. Under Review.
- [38] 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.
- [39] M. Perkowitz and O. Etzioni. Adaptive sites: Automatically learning from user access patterns. In *Proc. 6th Int. World Wide Web Conf., Santa Clara, California*, 1997.
- [40] J. M. Pierre. On the automated classification of web sites. *arXiv preprint cs/0102002*, 2001.

- [41] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, pages 98–105, 1998.
- [42] K. Schöfegger, C. Körner, P. Singer, and M. Granitzer. Learning user characteristics from social tagging behavior. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 207–212. ACM, 2012.
- [43] R.-N. Seamans et al. Technology shocks in multi-sided markets: The impact of craigslist on local newspapers. 2010.
- [44] M. Shmueli-Scheuer, H. Roitman, D. Carmel, Y. Mass, and D. Konopnicki. Extracting user profiles from large scale data. In *Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud*, page 4. ACM, 2010.
- [45] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. In *ACM SIGIR Forum*, volume 33, pages 6–12. ACM, 1999.
- [46] J. Stoyanovich, S. Amer-Yahia, C. Marlow, and C. Yu. Leveraging tagging to model user interests in del.icio.us. In *AAAI Spring Symposium: Social Information Processing*, pages 104–109, 2008.
- [47] A. Sun, M. A. Suryanto, and Y. Liu. Blog classification using tags: An empirical study. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 307–316. Springer, 2007.
- [48] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.