In Search of a Toyota Sienna: Prevalence and Identification of Auto Scam on Craigslist

Shirin Nilizadeh, Darya A. Orlova,*
Azadeh Nematzadeh, Minaxi Gupta, Apu C. Kapadia
Indiana University
{shirnili, dorlova, azadnema, minaxi, kapadia}@indiana.edu

Andrew Kalafut†
Grand Valley State University
kalafuta@gvsu.edu

Abstract

Craigslist ads are viewed by millions of Internet users each month, making it an attractive target for fraudsters and miscreants. Unsurprisingly, it has even been labeled a "cesspool of crime." In this paper, we take a first look at automobile scam on Craigslist. Focusing on the U.S. market, we find scammers are exploiting the fact that posting ads on Craigslist is free. They post a large number of ads for the same vehicle in many cities over a short period of time, either manually or by leveraging the easy availability of automatic ad posting software. Interestingly, scams often advertise relatively new vehicles of popular makes and list them at tempting prices. They extensively use special characters to attract attention and randomize ad body to escape automatic detection. Fortunately, our study finds many features distinguishing scam from good ads. Using these features we show that an SVM based classifier can differentiate between scam and trustworthy ads with 99% accuracy.

1 Introduction

Online auction and shopping websites such as Craigslist and eBay are used by millions of users to buy and sell a variety of goods and services worldwide. Craigslist alone operates in 570 cities in 50 countries. Its website publishes 80 million new classifieds every month which are viewed by over 50 million viewers over twenty billion times. Due to high traffic volume, `Craigslist.com` is ranked 36th among all websites worldwide and 7th among websites in the United States by Alexa [2]. Unsurprisingly, Internet fraudsters on the lookout for lucrative opportunities to find potential victims have taken note of these statistics. Recently, Craigslist was referred to as "a cesspool of crime" due to a high number of criminal activities [20]. A quick search on Google reveals many instances of fraud involving virtually every section of Craigslist including automobiles, tickets, housing, jobs, and services. The problem is so pervasive that Craigslist displays "personal safety tips" alongside all ad listings [7].

A few works have investigated the issue of scam on electronic auction sites such as eBay [5, 8] but scam on Craigslist is a relatively untouched topic to our knowledge. While one might be inclined to believe that fraud on all kinds of electronic market places might possess similar characteristics, there are a few important differences between auction sites such as eBay and Craigslist that suggest that an independent investigation on Craigslist is warranted. The first difference is that posting ads on Craigslist is free except in certain cases. (Craigslist charges job posts in San Francisco Bay Area at $75 per ad and in several other metropolitan areas at $25 per ad. Brokered apartment listings in New York and posts for therapeutic services are charged at $10.) This is not the case with eBay, where all but automobile posts

*The first two authors contributed equally.
†Work was done while the author was a Ph.D. student at Indiana University.

are free but sellers are charged a percentage of sale price upon the closing of an auction. This difference may translate into more scam on Craigslist. The second key difference between Craigslist and eBay is that eBay has a concept of *reputation* for both buyers and sellers while Craigslist does not have an equivalent concept. While reputation does not deter fraudsters from posting scam ads on eBay, they change the nature of fraud. For example, a common type of fraud reported on eBay is where fraudsters list inexpensive items through one account and then have many self-established accounts buy those items to create fake reputations which are later exploited for scam involving big ticket items, such as plasma TVs and laptops [9].

In this paper, we investigate automobile fraud on Craigslist. We explore the automobile section as against other Craigslist categories due to the first hand experience of one of the authors who was looking for a Toyota Sienna last year (hence the title of the paper). This experience gave us a window into the large world of auto scam, which prompted our deeper exploration. We focus on the automobile listings in the U.S. owing to our familiarity with it and the availability of commonly accepted vehicle prices through the Kelly Blue Book (KBB) website [12]. As a first step of our investigation, we collected 60 days worth of ads in the automobile section of Craigslist. Since Craigslist ads do not have a specific format, we developed various *parsing heuristics* based on extensive examination of many ads to extract useful features, such as vehicle make, model, year, price, mileage, phone numbers and email addresses. In order to characterize the nature of scam versus ads that could be trusted, we needed *ground truth*, which was challenging to establish. Fortunately, we were able to leverage two of Craigslist's unique functionalities toward this goal. First, it allows users to *flag* spam, miscategorized and prohibited ads. It also allows posters to *delete* their own ads. We found that flagged ads contained a high fraction of scam while ones deleted were overwhelmingly good. Using these as starting points, we iteratively sampled ads to find heuristics that help create sanitized ground truth. In the process, we found many unflagged ads belonging to campaigns for which some ads were already flagged. To find ads belonging to the same campaign, we used stylometric techniques to identify ads similar to flagged ads. The heuristics revealed that *scammers often post multiple ads to cast their net wide and about half of the ads belonging to known scam campaigns are not noticed by Craigslist, possibly victimizing many Craigslist visitors.*

Subsequently, we characterized scam and trustworthy ads. Our characterization revealed many interesting features of scam ads. For example, scammers tend to advertise popular makes of vehicles that are 3-5 years old and have low mileage (25-65K miles). They often ask for lower prices than their trustworthy counterparts, with $2,500-$6,500 being most common asking prices. Also, even though Craigslist provides an email address handle for interested parties to contact a seller without revealing seller's e-mail address, scammers tended to provide external e-mail addresses in ads to bypass Craigslist. Further, there were several unique features about the use of special characters, for example, the use of #, in the title of the ad and randomization of words in the body of the ad across ads belonging to the same campaign. While special characters may be useful in making a scam stand out, the motivation for randomization is likely for evading automatic comparisons for similarities. Also, many campaigns spanned all cities in our data sets and their ads were posted within a day, possibly indicating the use of automatic ad posting software.

We leverage features that differentiate scam and trustworthy ads to then train several classifiers using supervised machine-learning. We find a support vector machine (SVM) based classifier performs best and is able to achieve an accuracy of 99% for differentiating between trustworthy ads and scam ads. Our classifier also provides 99% precision and recall for these categories. Thus we show that supervised learning techniques can be highly effective at automatically detecting scams on Craigslist.

2 Data Collection Methodology

We focus on Craigslist automobile classifieds in the U.S., a market we are more familiar with and because it allows us to estimate vehicle prices based on data from Kelly Blue Book (KBB) [12]. In the U.S., the classifieds are listed under 413 cities. Of these, we focus on the 30 largest metropolitan areas as identified by Craigslist. Craigslist publishes automobile classifieds in the *cars+trucks* section. Within this section, there are two categories: *by-owner* and *by-dealer*. We choose to investigate the by-owner

classifieds upon noticing that scam seemed easier to detect in that section, leaving the investigation of the by-dealer section to future work.

2.1 Data collection

For our analysis, we first needed to create a corpus of classified ads. To this end, we developed a Perl script that crawls automobile classifieds listed under the by-owner section. Our data collection started on 11/19/2010 and ran daily for a period of 60 days. Notice that our data collection started before the Thanksgiving break and included the peak Christmas shopping period, possibly capturing scam targeting the holiday period. Further, the post new year period was intended to capture the off-peak scam.

We collected data in two phases. In the first phase, our crawler collected the HTML representation of all ads (subsequently referred to as "ads") in a breadth-first fashion, for all of the 30 metropolitan areas in parallel. Each time it is run, it collects all available ads in a reverse chronological order, following Craigslist's indexing order. Since Craigslist keeps classifieds for 7 days in 10 of the 30 metropolitan areas and 30 days in the rest of the cities, our crawler collects 7–30 days of ads in its very first run. On subsequent runs, the crawler captures only newer ads and any changes to older ads. We run the crawler daily for the entire 60-day period.

Craigslist assigns a unique identifier to each ad, which facilitates detecting changes. We store each ad when we encounter it the first time or when it is found to have changed in subsequent crawls. Specifically, there are two types of changes that are interesting. The first is when an ad is *flagged for removal*, which can happen for multiple reasons, including when an ad is scam. This change is reflected in the form of a generic "flagged for removal" message from Craigslist. The second reason an ad can change is when it is deleted by its poster. This change is reflected in the form of a standard "deleted by author" message from Craigslist. Based on these two types of changes, we label the ads as either *flagged* or *deleted_by_author*. The marked ads are later used in establishing ground truth for scam and good ads as described in Section 3. Notice that if an ad is edited, Craigslist creates a new identifier (and hence URL) for it and our crawler will consider the changed ad to be a new ad.

Craigslist frequently refreshes its listing index and removes all flagged and deleted ads, and running the crawler daily might miss some flagged and deleted ads. Thus, in the second phase of our data collection, we run a review crawler which leverages the fact that though Craigslist does not display flagged or deleted ads to its viewers, the URLs corresponding to those are still accessible. Accordingly, the review crawler simply revisits each unique URL and records if the corresponding ad was flagged or deleted. We ran the review crawler on two occasions, once on 1/21/2011 and again on 2/04/2011. The first review found that most ads that were flagged were 2 weeks or older. This suggests that it takes some time for users to recognize a fraudulent ad and flag it. For this reason the second review was conducted on 1/21/2011 with the intention of capturing more flagged and deleted ads from early January.

2.2 Parsing and metadata extraction

In order to be able to study various characteristics of good and scam ads, we parse all ads for the make, model, year, price, and mileage of the vehicle, presence of non-Craigslist URLs, date of the original post, phone number, e-mail address, user-created ad text, images, title of the ad. Parsing presented some challenges since there are many ways for sellers to represent information. For example, "10000 miles" can be expressed as "10K mi", "mileage: 10,000", "10xxx" and so on. Based on several observed examples we developed a regular expression which checks for "miles", "mileage", "mi", "mile" before or after a number or a number followed by a "k" or "xxx", case-insensitive. In order to extract makes and models from ads, we compiled lists of all makes and models available at KBB. A limitation of this approach was that we could only extract makes and models for cars made after 1990 because KBB does not have older makes and models available. However, this approach allowed us to identify makes and models in 86% of deleted and 78% of flagged ads, indicating that the loss in information due to this strategy is probably minimal. We corrected for 4 common misspellings in the make and model names that were found via manual examination of a sample of ads: "Corla", "Acura", "Chevy", "Acord", but for the most part misspellings were not corrected as we did not see any other obvious patterns in our manual inspection. Year was identified as a 4-digit number between 1990 and 2012 or a 2-digit number

which was then converted to what might be a year between 1990 and 2012. If the number which was parsed out as year is not within the range of 1990–2012, we consider the year to be invalid.

Many ads list contact information in the form of phone numbers, email addresses, or postal addresses. In our experience, postal addresses were the least common. Also, we observed a huge variation in formats in which postal addresses were listed. Due to these reasons, we only focused on phone numbers and email addresses in the context of contact information. To find phone numbers we search for strings of 3 numbers, 3 numbers, 2 numbers, and 2 numbers, separated by some combination of spaces, dashes, dots, and parentheses. A phone number with only 7 digits would not match this definition, as would a phone number with some digits spelled out (e.g. 234-4five9-0two01). Emails are found with a regular expression matching any number of letters, digits, dashes, dots and underscores followed by an symbol and followed again by any number of letters, digits, dashes, dots and underscores, separated by at least one dot. E-mails written out in words are not captured through this technique, however (e.g. mickey-mouse AT disneyland DOT org). Most Craigslist ads contain an e-mail address in the Craigslist domain. We refer to it as the *Craigslist handle*. Extracting this handle is straightforward and we extract it for all ads that contain it.

An overview of data collected is given in Table 1.

Type of advertisement	Number of advertisements
Unique ads	2,424,092
Flagged for removal	42,185
Deleted by author	612,255

Table 1: Overview of the types of advertisements collected. The dates of the ads covered a 3.5-month period.

2.3 Limitations of data collection

Our data set is limited in scope in that we focus only on by-owner scam in metropolitan areas and that we do not investigate scam outside the U.S. While the first two choices were governed primarily by the need to avoid our crawler from being blacklisted by Craigslist, the last choice was made because it allowed us to estimate the value of an automobile. However, a careful examination this limited data set allowed us to learn how scammers operate on Craigslist. Whether our findings apply to international automobile scam or scam in general remains an area of future investigation.

3 Establishing Ground Truth

To characterize known scam and known good ads and extract features for automated detection through classification, we needed to establish ground truth for each type in our corpus. Manually labeling ads as *scam* or *good* in our corpus would be prohibitive, so we established ground truth in two steps: first we generated `flagged` and `deleted` data sets from out corpus, and then we filtered these data sets based on various criteria so that these data sets represented scam and good ads respectively.

3.1 Generating data sets

Our data contained two cues to help establish ground truth. First, Craigslist users flag ads in three categories: *miscategorized* (misplaced or in wrong category), *prohibited* (violates Craigslist Terms of Use or other posted guidelines) or *spam/overpost* (posted many times, in multiple cities or categories or is too commercial). Craigslist removes ads receiving a sufficient number of negative flags. Once removed, retrieval of URLs corresponding to flagged ads displays the "flagged for removal" message, which we leverage to find potential scam ads. Note that this message is generic, in that it does not tell if the ad was miscategorized, prohibited, or spam/overpost. So, it cannot be used to find scam ads as is. However, it gives us a starting data set that we filter to create a corpus of scam. We refer to this data set as the `flagged` data set subsequently. The second cue in our data was the "deleted by author" message Craigslist displays when posters delete their ads. While scammers are unlikely to delete their own ads because they would want them to stay up as long as possible to catch more victims, legitimate sellers may choose to take their ads down upon selling their vehicles to curtail requests from prospective buyers.

An examination of a random sample of deleted ads suggested that these ads were mostly not scam (but did include spam as we discuss later). Again, we use this data set, referred to as the `deleted` data set subsequently, as a starting data set that we filter to generate a corpus of good ads (see Section 3.2).

3.2 Filtering the data sets

Establishing ground truth starting from "flagged for removal" ads and "deleted by author" ads posed multiple challenges. In addition to the fact that flagged ads contained miscategorized, prohibited and spam ads along with scam, we found many good ads are also sometimes flagged for personal reasons, such as by competing sellers or even by buyers who wish to make items unavailable to other buyers. In fact, a few of the automatic ad posting programs we encountered on the Web (e.g., CladGenius, Craigslist Bot Pro, Craigslist Classified Ad Posting Utility, and ClassifiedAutomation.com) had a provision to flag ads. Further, many deleted ads fell in the spam category and many were selling car parts or even things that had nothing to do with cars.

To account for these factors, we applied a two-step technique to both `flagged` and `deleted` data sets. In the first step, we find ads with a valid make, model and year in the flagged data set and ads with valid make or model in the deleted data set. The subtle difference in filtering the two data sets was due to the observation that scam almost always included a valid make, model and year, perhaps to maximize the chances of the ad being found in searches while good ads tended to skip either make or model on many occasions. Additionally, since scammers tend to advertise newer vehicles at tempting prices, we ensured that ads listing vehicles built in year 2000 or later at "tempting prices" (as described below) were the only ones retained in the flagged data set. Notice that some ads in the deleted data set also had tempting prices but in most cases, their low price could be justified because the vehicle was in an accident or had a rebuilt title. This criterion was crucial to sanitizing the flagged data set since without it, there were many misclassifications. To determine tempting price for a given make, model and year combination, we retrieved the KBB price for each vehicle in good condition with either the listed mileage or an estimate of 15,000 miles per year. We use private party values which are a suggestion for fair price of a vehicle if individuals transacted among themselves. These values are lesser than when vehicles are purchased from an automobile dealer and more than when they are sold to a dealer. Further, upon experimenting with various thresholds, we defined price to be tempting when a vehicle was listed for less than 75% of the KBB price. Tables 2 and 3 show the impact of this filtering step on the `flagged` and `deleted` data sets respectively.

In the second step, we iteratively sampled 150 ads from each category and identified keywords in ads that were incorrectly categorized. We then removed all ads containing those keywords from each data set. To minimize removing non spam ads, we sampled ads that got removed from each set at each iteration and adjust our list of keywords. This process ended until our random samples of each data set contained no spam.

The keywords we used for filtering fell in two categories. In the first category are keywords that identified spam. This included posts where other buyers were warning about spam. This list had the following 25 keywords: "warning", "scam", "fraud", "be aware", "beware", "too good to be true", "alert", "stolen", "fake", "spam", "junk", "unwanted", "repossessed", "rebuilt title", "rent" , "trashed" , "smashed", "wreck", "free", "we buy", "we pay","any condition", "get paid", "trade", "everything must go". The second category of keywords included miscategorized ads, including those where car parts, things other than cars were sold or where dealers were advertising when their ads are supposed to be in the dealer portion of the cars+trucks section. This list had the following four 4 keywords: "wheels", "parts", "part car", "dealer". Additionally, we use the criterion that the quoted price on the ad had to be less than $500 since most ads for car parts were for a lesser amount. Note that we applied these keywords to the entire body of ads, including titles.

Table 2 contains the result of filtering. As it shows, filtering removed 94% of the ads from the flagged data set, leaving us with 2.5K scam samples. We refer to the resulting data set as scam data set subsequently in this paper. Similarly, filtering the deleted data set left us with 71% (437K) ads. We refer to this data set as the `trustworthy` data set subsequently. Notice that a large portion of the data we collected is not labeled as trustworthy or scam. We discuss how we use that data set, `unlabeled`,

subsequently to expand on our scam corpus.

	flagged	%
Total ads	42,185	100
Valid make, model & year	12,469	29.56
Year ¿ 1999	9,279	22.00
Price found in ad	8,019	19.01
KBB price retrieved	6,374	15.11
Temptingly priced	2,907	6.89
No spam keywords	2,531	6.00
Size of scam data set	**2,531**	6.00

Table 2: Effect of filtering steps on the flagged data set. The resultant data set is called scam.

	deleted	%
Total ads	612,255	100
Valid make or model	526,863	86.05
No spam keywords	437,238	71.41
Size of trustworthy data set	**437,238**	71.41

Table 3: Effect of filtering steps on the deleted data set. The resultant data set is called trustworthy.

4 Does Scam Go Unnoticed?

Our personal experience with scam campaigns on Craigslist was that in a typical case, a campaign involved multiple ads. In fact, a search on the Web revealed that many tools were available to automate posting of ads on Craigslist. These automators featured a variety of options and their prices ranged from free to $100. YouTube demos of such software revealed that given these automators could simultaneously post many ads while varying ad titles, body, and prices. Often, they inserted extraneous characters, such as $, @, or !, or adjectives describing the condition of the car in the title. Also, insertion of extra words in the ad body and randomization or words was a common feature. The price variation they introduced was within a few percent points of the original price selected by the poster. In fact, a few of the software would also automatically flag existing ads for same or similar vehicles. Given this information, we wondered if ads belonging to campaigns we found in the scam data set were present in the unlabeled data set because they were either not flagged or the extent of flagging was insufficient for Craigslist to remove them.

As a concrete example, the ads in Figure 1 ads clearly belong to the same campaign but only the first one was present in the scam data set. Notice that these two ads have similar styles. Their titles are identical, the text is similar but scrambled. The ads are posted on the same day with identical prices. Further, in spite of the similarity of their content, their posting cities are far enough from each other that they are unlikely to be from one legitimate poster wanting to expand to neighboring cities.

Subsequently in this paper, we refer to ads belonging to the same campaign as *sister ads* . Also, we define a *campaign* as all sister ads for vehicles with same make, model and year and similar price. We make this choice since none of the posting automators we examined vary make, model or year of the vehicle. Also, they all vary prices within a small percentage of the original price specified by the poster. Notice that if a scammer posts multiple ads for different vehicles, each resulting in multiple ads in various cities, our definition will consider them as separate campaigns.

4.1 Finding potential sister ads

To expand our scam corpus we took ads in the scam data set and looked for sister ads in the unlabeled data set. We also tested the trustworthy data set for the presence of any sister ads and found a small number there, indicating that the iterative filtering and sampling of 50 ads out of the 612K ads in the trustworthy data set left some scam undetected due to the large size of this data set. We eliminated that

2007 Dodge Durango 4x4 Very Nice - $5720 (miami)

Date: 2011-04-07, 12:29PM EDT
Reply to: sale-gfqqp-2310224458@craigslist.org [Errors when replying to ads?]

remote hood/fuel door releases power mirrors, power door locks, power windows, seat, cloth seats, security system,adjustable headrests a
lock brakes console, center armrest, split bench

- Location: miami
- it's NOT ok to contact this poster with services or other commercial interests

PostingID: 2310224458

(a) Flagged scam ad

2007 Dodge Durango 4x4 Very Nice - $5720 (sacramento)

Date: 2011-04-07, 9:29AM PDT
Reply to: sale-xdwdd-2309707068@craigslist.org [Errors when replying to ads?]

remote hood/fuel door releases power mirrors, power door locks, power windows, console, center armrest, split bench seat, cloth seats, security system,adjustable headrests anti-lock brakes

- Location: sacramento
- it's NOT ok to contact this poster with services or other commercial interests

PostingID: 2309707068

(b) Unlabeled scam ad

Figure 1: Two scam ads from the same campaign.

scam to further filter the trustworthy data set. The steps of the algorithm we used to identify sister ads are the following:

Create templates for campaigns We began by identifying *templates* of individual scam campaigns present in the scam data set. We define a template as the collection of scam ads with same make, model and year and similar price where the similarity constraint forces all ads in a campaign to be within $1000. We choose $1000 as the range because all sister ads we sampled fell within that range.

Prune campaigns spanning less than 5 cities If campaigns resulting from the previous step span less than 5 cities, we remove them from consideration since a sampling showed that many of the ones in less than 5 cities were often postings by zealous sellers in neighboring large cities. Further, campaigns using automatic posting software are likely to be spread to a large number of cities.

Identify potential sisters In this step, we identify potential sister ads for each template by examining the unlabeled and trustworthy data sets. Toward this goal, we apply exactly the criterion we used to define campaigns, that is, ads with the same make, model and year and similar price to known scam are considered sisters. For the price, we take the minimum and maximum prices in each template and allow for 5% variation on either side.

4.2 Filtering sister ads

The potential sister ads we identified in the previous step contained many false positives because legitimate ads for the same make, model and year could have a tempting list price just like scam for myriad reasons including difference in trim, high mileage or because the vehicle was in an accident. We prune ads falsely categorized as sisters in the subsequent steps. Two ads from different cities with the same make-model-year are highlighted in the Appendix in Figure 12 as examples of non sisters.

Campaigns either have images or not We noticed that ads within a campaign either all had images or none of them had images. (The number of images sometimes varied across ads of the same campaign.) Using this observation, we pruned campaign templates. We also pruned sister ads without images that fell in campaigns where all ads had images and vice versa.

Sister ads share contact information, if any We found that if a scam listed a phone number, all ads from the same campaign listed the same phone number. Similarly, if a scam listed an email address, all ads from the same campaign listed an email address, though not necessarily the same email address. Since several automatic posting software allowed posters to specify a group of email addresses that they would rotate through in posts, this test simply tries to capture the behavior of automators. To account for these observations, we checked if ads in a template contained contact information. If so, we kept a newly identified sister if it contained the same phone number as the template or an email address, as the case may be. Otherwise, we subjected it to the following tests to see if the sister ads matched the campaign based on other criteria.

Sister ads share mileage Similar to contact information, all sister ads belonging to the same campaign either contained the same mileage, if any. Based on this observation, if template ads contained mileage, we kept those newly identified sisters that listed the same mileage. The rest were subjected to the following tests.

Sister ads have similar titles We noticed that titles of ads in the same scam campaign varied along three dimensions: presence of punctuation (e.g. $\sim\sim 2004HondaOdyssey \sim\sim Loaded \sim\sim RunsGreat \sim\sim$), permutations of word order, or presence of adjectives. Specifically, many scam campaigns contained combinations of adjectives such as “mint”, “awesome”, “excellent”, “fantastic”, “nice”, “wonderful”, and “clean” in the title of the ad. In fact, ads within a campaign often saw these adjectives repeated, which is why we test if ads in a template contain any in the title. If so, we consider a newly identified ad to be a sister if it shares any other words in the title in addition to make, model, year and also contains the same adjective as at least one ad in the template. Also, to avoid the issue of false positives, we further imposed the restriction that a newly identified sister also had to have the exact same price as an ad in the template.

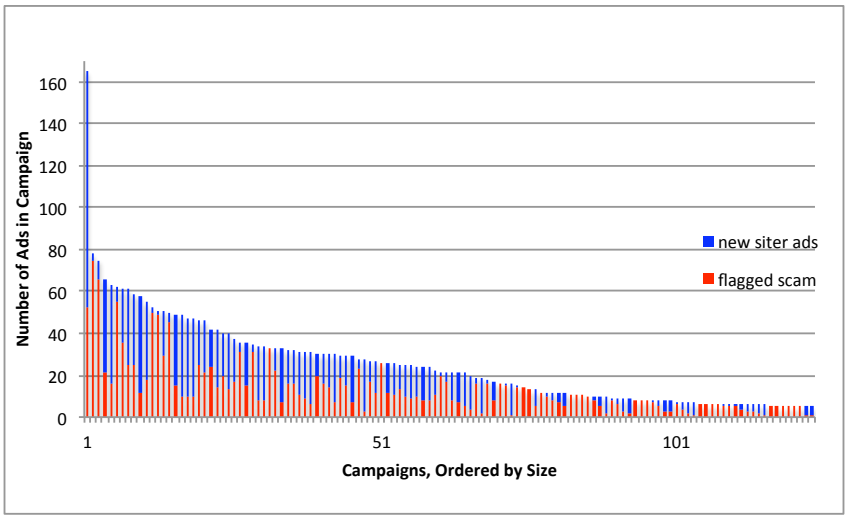

Figure 2: Largest 106 scam campaigns, ordered by size

Sister ads have similar bodies Automatic ad posting programs frequently alter the body of the original ad to create different-looking ad text. Two techniques used for this purpose are: reordering of the same set of words within each ad text and creating ads by concatenating a randomized subset of phrases from a single pool of phrases. To find simple word order changes, we order all words in each ad alphabetically, remove duplicate words, and simply compare the resulting lists. For several ads created from a single pool of words we extract a set of all words present in scam ads from the same campaign. Next, we check how many of the words from that pool are in a potential sister ad. In both cases only words of length 3–20 characters were used since words with one or two characters are often not real words or are stop words which are commonly eliminated from natural language text during stylometric analysis. In our case, non-words are things like "lx", "se", which usually represent vehicle trim. Also, words over 20 characters long are usually not words either. Filtering them out is important because scammers sometimes add strings of letters to the end of an ad, presumably to make the content appear different. Such strings are always different but do not contribute to readable text.

We calculated what percentage of words in a potential sister ad were also found in the flagged word pool. If the percentage was 75% or above, the ad was considered part of the campaign. If the match was less than 25%, it was not considered part of the campaign. For matches between 25% and 75%, we only considered the newly identified sister to be part of the campaign if its price matched the price of a known scam from the same campaign. Manual examination of sister candidates showed that in cases were the words in ad text were 25-75% similar to known flagged ads, identical price was a good indicator that the potential sister did indeed belong to the campaign.

Table 4 shows the total number of templates present in the scam data set. 1,581 (62%) of scams belonged to one of the 761 templates. We started with 2,517 potential sister ads. Upon confirming or eliminating sister ads based on the above algorithm, we found 1731 new sister ads. We observe *only 48% percent of the scam ads from known campaigns were flagged on Craigslist. The rest stayed posted, potentially luring users looking for vehicles*. Figure 2 shows how the new sister ads expanded the largest 106 campaigns consisting of more than five scam ads. Specifically, for the largest campaign, only 52 ads were flagged. Our algorithm found an additional 113, expanding the size of that campaign to 165 scam ads.

5 Scam Characterization

We learnt several characteristics of scam and trustworthy ads while filtering the `flagged` and `deleted` data sets to obtain ground truth for scam and trustworthy ads as described in Section 3. Specifically, we learnt that scam tended to advertise newer cars with valid make, model and year at tempting prices. Similarly, in the process of finding unflagged sister ads belonging to known scam cam-

Type of ads	# ads
Total ads in scam data set	2,531
Campaign does not span 5 or more cities	875
Did not pass "campaigns either have images or not" test	36
Total scam ads passing the above tests	1581
Total campaign *templates*	761
Potential sister ads in unlabeled data set	2,109
Potential sister ads in trustworthy data set	408
Total potential sister ads	2,517
Sister ads *confirmed* via shared contact information	260
Sister ads *confirmed* via shared mileage	468
Sister ads *confirmed* via title similarity	625
Sister ads *confirmed* via body similarity	378
Total *confirmed* sister ads	1,731
Sister ads *eliminated* due to "campaigns either have images or not" test	167
Sister ads *eliminated* because they did not match any criterion	619
Total sister ads *eliminated*	786

Table 4: Confirming sister ads to known scam templates

paigns in Section 4, we learnt several characteristics of scam campaigns. First, they often spanned more than 5 cities. All ads in a campaign either all contained images or none did. Also, sister ads shared contact information and mileage. Even the body and title of sister ads had significant stylometric similarity. In this Section, we further characterize scam in order to find differentiating characteristics with respect to trustworthy ads. We characterize scam ads individually as well as at the campaign granularity in Sections 5.1 and 5.2 respectively. For scam, we include sister ads in the `scam` data set. For trustworthy ads, we use the `trustworthy` data set (with sisters removed).

5.1 Characteristics of Individual Scams

Time of year We begin by examining if there is any difference in scam around the Christmas holiday season versus after the new year. Figure 3 shows the number of ads posted around Thanksgiving through the end of the data collection period in January 2011. In comparison to the first few weeks in January, both scam and trustworthy ads increase in numbers until a dip around Christmas time, suggesting that scammers are not particularly active around the holiday period.

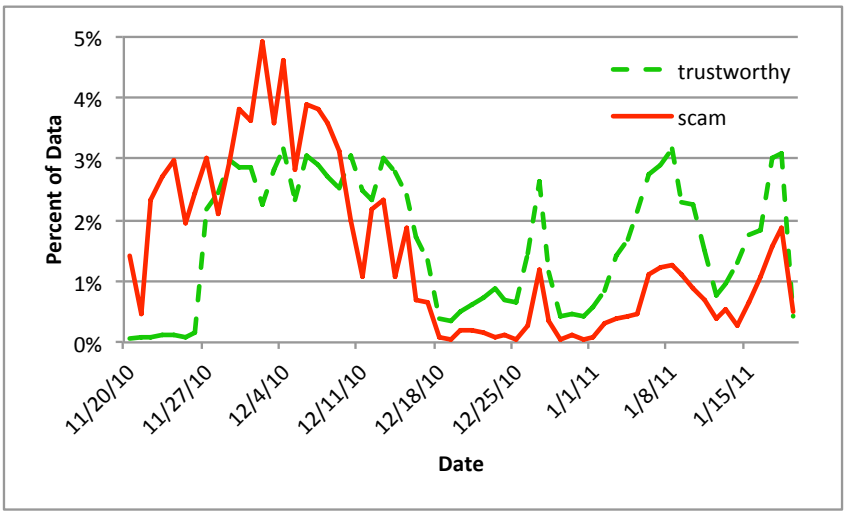

Figure 3: Number of ads on each day

Vehicle makes Here, we look at the makes of vehicles advertised in scam versus those in trustworthy ads (Figure 4). All ads in the scam data set and 92.7% in the trustworthy data set had makes listed. Note that Dodge, Ford, Honda, Nissan and Toyota are most popular makes in trustworthy ads as well as scam.

The only exception is Chevrolet, which is more popular in scam ads. We also observe Honda, Ford, BMW, Dodge, and Jeep are more prevalent in trustworthy data.

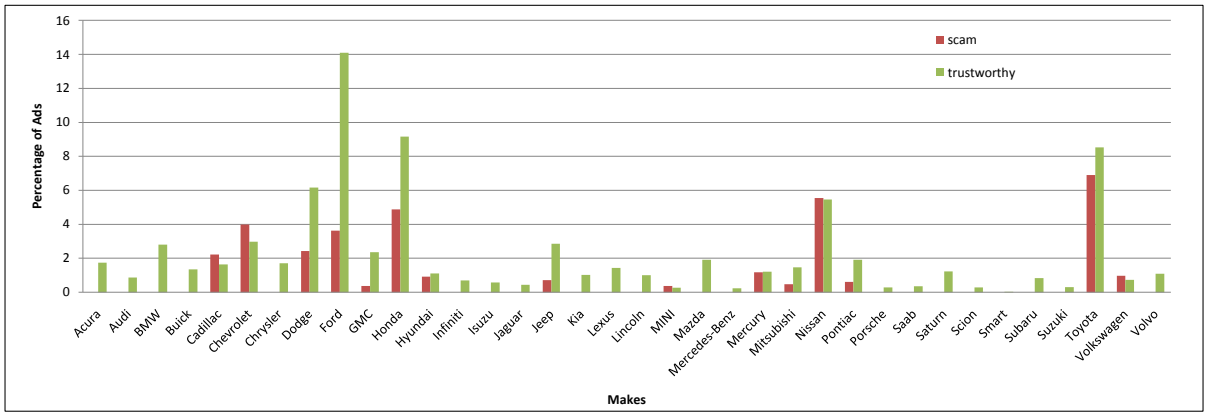

Figure 4: Distribution of vehicle makes

Vehicle year 82.9% of ads in the trustworthy data set listed vehicle year between 1990 and 2012. 45.8% of these years were greater than 1999. By comparison, all ads in the scam data set had a year great than 1999 because of our selection criteria. We note that the frequency of trustworthy ads with year greater than 2000 declines steadily, whereas frequency of scam reaches a peak at 2003 and then drops off gradually. It thus appears that scammers prefer to lure people with vehicles that are about 5 years old.

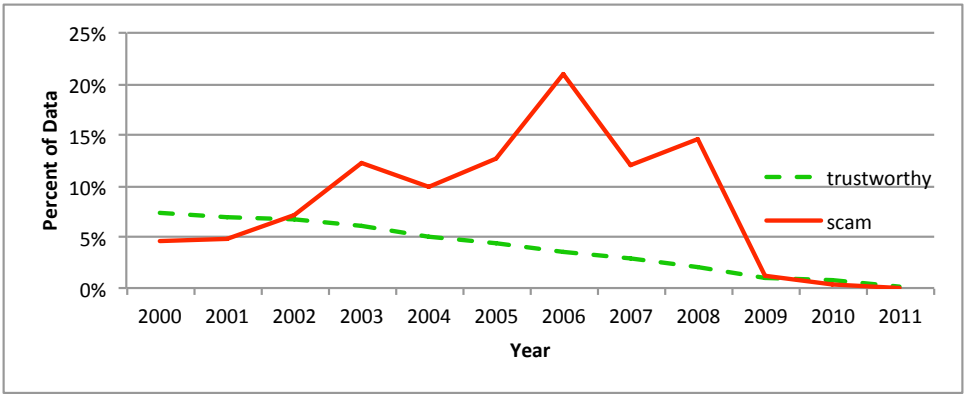

Figure 5: Distribution of vehicles by year

List price Price was listed for 94.5% of trustworthy ads and for all scam ads. Figure 6 shows list prices rounded up to multiples of $500. Note that because we used price as a filtering criterion to create the scam data set, price in scam is always 75% or less of the KBB price. Nonetheless, the Figure shows that vehicles advertised in scam often have prices between $2,500 and $6,500 while prices in trustworthy ads have a much wider range. This suggests the scammers find it lucrative to target scams in that range of price.

Mileage 38.9% of ads in the scam data set and 51.5% in the trustworthy data set contained mileage. Here, we plot the distribution of quoted mileages in Figure 7. Note that while mileages for trustworthy ads have a large spectrum, a large fraction of scams listed low mileage, between 25K to 65K miles.

Contact information As shown in Table 5, most ads contained either e-mail addresses or phone numbers. Craigslist e-mail addresses were more prevalent in trustworthy ads than scam. The difference was more pronounced for external e-mail addresses where trustworthy ads contained an e-mail address belonging to non Craigslist domains far less often, 1.2% as against 17% with scam ads. Scam tended

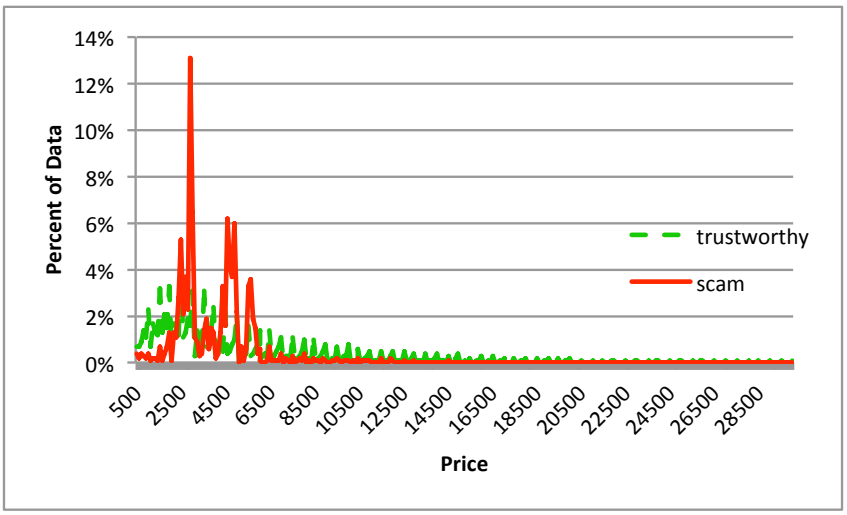

Figure 6: Distribution of prices

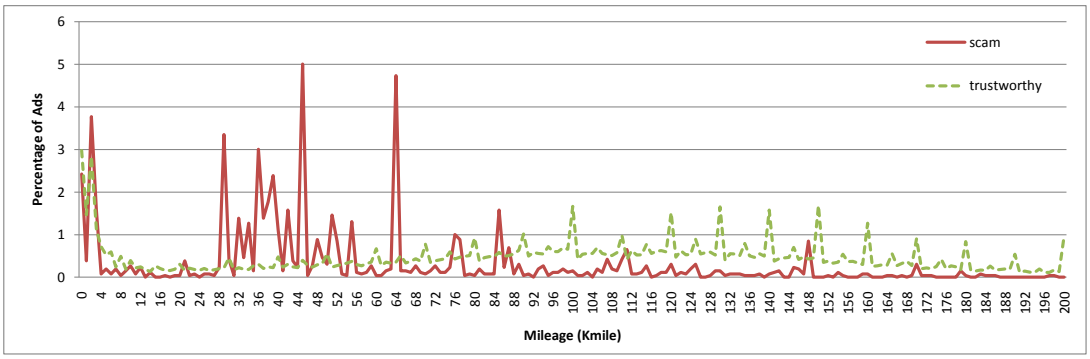

Figure 7: Distribution of mileage in ads

to have phone numbers far less often than trustworthy ads, which Craigslist recommends its users look at as a criterion to judge if an ad is trustworthy or not. Both trustworthy and scam ads sometimes contained syntax designed to avoid being collected by a crawler. Trustworthy most commonly listed phone numbers with digits spelled out or an e-mail address with ”at” instead of ”@” and ”dot” instead of ”.”. Scam occasionally used images with a phone number or e-mail contact embedded in the image.

	trustworthy%	scam%
E-mail or phone	97.9%	90.4%
Craigslist e-mail handle	83.7%	69.6%
External e-mail	1.2%	17.3%
Phone number	80.2%	14.2%

Table 5: Contact information in trustworthy and scam data sets

Images Many ads, 89.8% of trustworthy and 95.6% of scam contained images. Craigslist only hosts up to four images per ad, but many ads across categories contain many more images linked from external websites. Such linking is convenient for a scammer because he or she can post a single picture and link to it in many ads. However, many honest users employed the same techniques, so it is is not predictive of scam.

External links External URLs were not very common in both trustworthy and scam data sets, with links found in 16.9% and 11.6% of trustworthy and scam ads, respectively. Thus, there are no major differences between trustworthy and scam ads in this dimension, implying that presence of external links is not indicative of scam.

Title of an ad Here, we investigated if titles of scams differed from those of trustworthy ads. We looked at the number of words, the distribution of their lengths and number and distribution of special characters in ad titles. While the words and the distribution of their lengths did not differ, the number and distribution of special characters in scams differed. Specifically, scam titles had a clear peak at six special characters. Further, scam ads had a clear preference for ’#’, ’!’, ’*’, and ’.’ in the title as shown in Figure 8. Of these, ’!’, ’*’, and ’.’ were popular in trustworthy ads as well but ’#’ was more preferred by scammers.

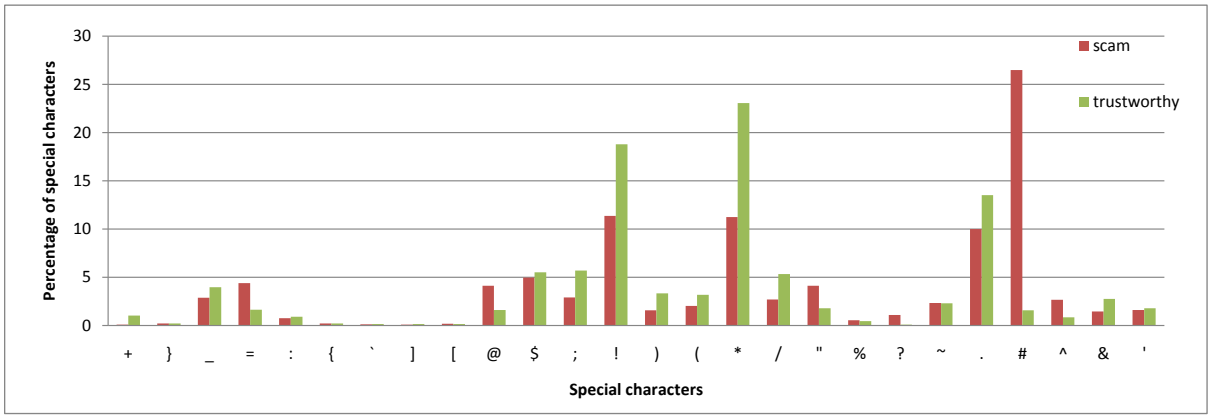

Figure 8: Distribution of special characters in ad titles

Body of ads Next, we looked at the number and distribution of sentence and word lengths and special characters in ad bodies. We found that scam and trustworthy ads differed little in terms of number and distribution of sentences. However, scam ads tended to have 35 words or between 55-80 words more often than trustworthy ads (see Figure 9). Also, scams often had 6-12 special characters more often than trustworthy ads though their distribution was not particularly interesting.

Identifying features based on the frequency of words (unigrams) or pairs of words (bigrams) is a typical approach in Text Mining. However classifying based on textual features may not be effective

enough in this context where scammers may imitate the writing patterns of normal users. Since such text extraction techniques are standard for classifying text, we present them in the next section. Our classifier will make use of those standard features as well as the features identified in this section.

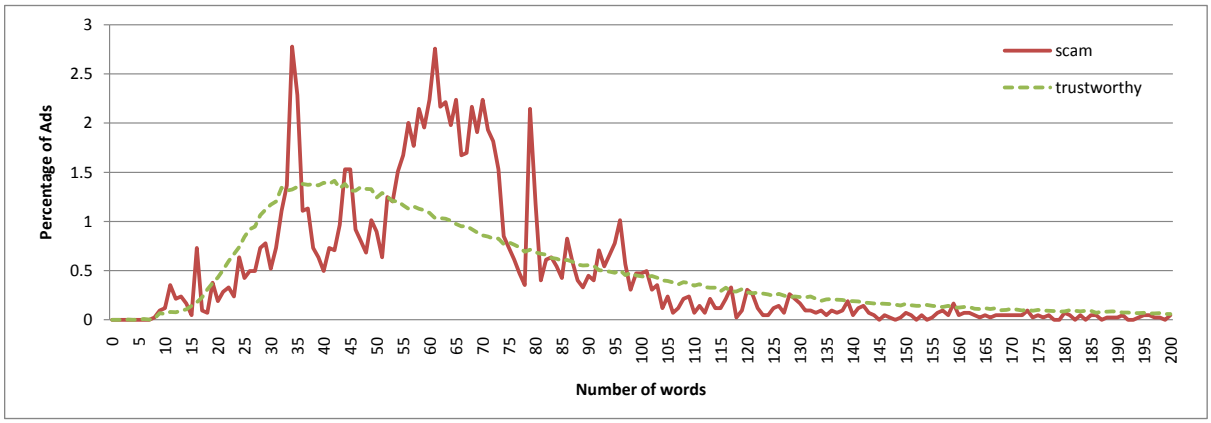

Figure 9: Number of words in ad body

5.2 Characteristics of scam campaigns

We characterize scam campaigns in this Section. Note that since we do not have a notion of campaigns in the trustworthy data set, the characterization cannot serve to highlight any differences.

As shown in Figure 10, scammers usually attempt to cover a large geographic area. In our data 53.8% of campaigns posted to 10 or more cities and 8% posted to over 20 out of the 30 cities for which we collected data.

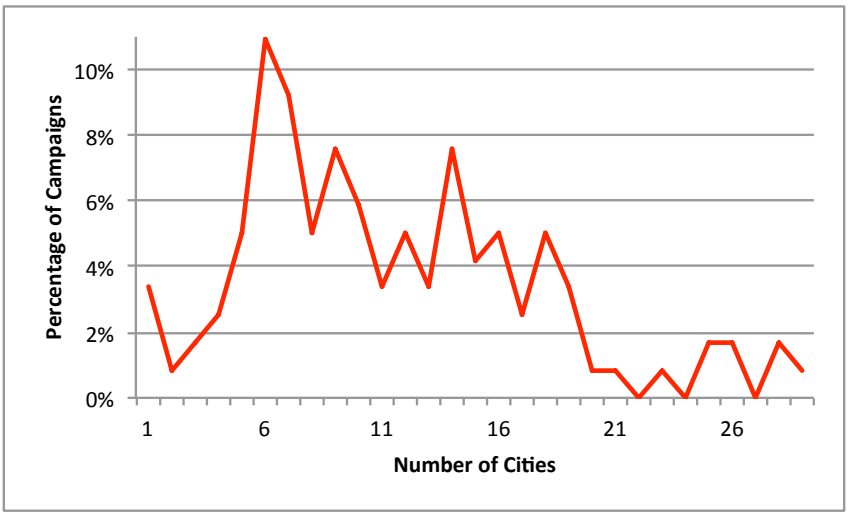

Figure 10: Number of cities in individual campaigns in the scam data set

Next, we looked at the duration over which ads within the same campaign were posted. Campaign duration ranged broadly from one day for 8 fastest campaigns to 108 days for the longest campaign. However, 57.6% of the campaigns lasted 30 days or less. Figure ref(fig:duration-distr) shows how campaign duration distribution.

6 Classification

In this section, we first explain the pre-processing transformations that were performed on the corpus before training the classifiers. We feed the classifiers with textual features and *entities*. We use bigram (word pairs) and SP (short window pairs, a special kind of bigram) features as textual features. Entities

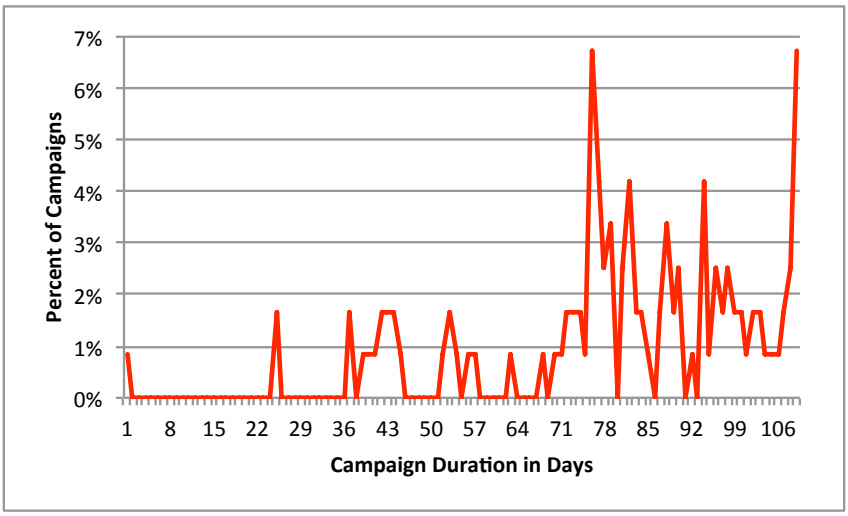

Figure 11: Duration of ads within individual campaigns

are any possible distinct characteristics of trustworthy ads and scams. We extract entities based on our analysis presented in Section 5. Finally, we present and compare the performance results of running several classification algorithms on these features.

6.1 Textual feature selection

Classification machine learning algorithms operate on numerical quantities as inputs. A labeled example is often a vector of numeric attribute values with one or more attached labels. The conversion of text records to vectors of numeric attributes is a multi-staged process of feature construction, which employs several transformations. Each unique word constitutes an attribute. The number of occurrences of a word in a record (frequency of occurrence) is the attribute's value for that document. Records are therefore represented as vectors of numeric attributes where each attribute value is the frequency of occurrence of a distinct term.

Often, simply extracting all unique terms is not the most effective way to acquire attributes for adequate classification. Some types of words, or series of words, may be preferable for learning. For example, often nouns or noun phrases are preferred. Also, common words like "and" and "the" are often filtered out to improve performance. All of these transformations should be performed before any learning takes place.

We employed three steps of transformations in the pre-processing phase including: 1) removing stop words i.e., the most common, short function words, such as "the", "is", and "which", 2) filtering out words with less than two characters, 3) stemming words, i.e., removing the common morphological and inflectional endings from words in English using the Porter Stemming Algorithm [19].

Features can be unigrams, bigrams or short-window pairs (SP). Unigram features are words that occur more than a given number of times in the corpus. Bigrams are frequent ordered word pairs. Short-window pairs are bigrams that are selected over the top 1000 unigrams of a document instead of all unigrams.

We ranked extracted features with S scores [14]. Top features should be a good discriminator features that are either more frequent in a positive or a negative class. The two classes are scam and untrustworthy in our case. The S score measures the difference between the probabilities of occurrence in the positive and negative training set documents and it is calculated as follows for each word w_i.

$$S(w_i) = |P_{TN}(w_i) - P_{TN}(w_i)|$$

Where $P_{TP}(w_i)$ and $P_{TN}(w_i)$ are:

$$P_{TN}(w) = \frac{|\{d|w \in d\}|}{|TN|}, d \in TN$$

$$P_{TP}(w) = \frac{|\{d|w \in d\}|}{|TP|}, d \in TP$$

Entity	Description	Value
Mileage in ads	If (35¡mileage¡40) or it is equal to 29, 45 or 85, it is more scammy	Binary
Special chars in title	Number of occurrence of sharp (#), exclamation (!), and star(*)	Count
Special chars and punctuation count in title	If number of special characters and punctuation is equal to 6	binary
Contains email	If the ad contains email address	binary
Contains phone	If the ad contains phone number	binary
Contains Craigslist handle	If there is Craigslist handle in the ad	binary
Contains image	If the ad contains an image	binary
Word count in body	If number of words is less than 53. Also, if greater than 75	binary
Make	If ad contains Honda, Ford, BMW, Dodge, Jeep since their count was varied in scams and trustworthy ads	binary

Table 6: Entities that we used in classification along with textual features

The S-measure helps us pick the features that are good discriminators. We found that SP and bigram features would be more discriminative than unigrams, and thus do not use unigrams to train our classifier. The top 20 unigram, SP and bigram features are listed in Table 9. The positive/negative probability and S score of the top 10 SP features and bigrams are listed in Table 10 in the Appendix.

6.2 Entity Recognition

We use several features we identified from our data set as described in Section 5. These entities are discriminatory features that can separate scam and trustworthy ads. For each ad, we calculated the entity count (e.g., "number of miles") or the binary value (e.g., "contains image?"). The extracted entities are listed in Table 6.

6.3 Classification

We studied the performance of several non-linear classification algorithms to see which one fits best for scam data analysis. We chose classifiers known for their high accuracy such as Sequential minimal optimization (SMO) [18] and Bagging with Random forest [3]. SMO is an algorithm for solving the optimization problem of support vector machines (SVMs). The main idea of an SVM is to construct a nonlinear kernel function to map the data from the input space into a possibly high-dimensional feature space and then generalize the optimal hyper-plane with maximum margin between the two classes. Training an SVM requires solving a large quadratic programming problem which is often intractable. SMO is an iterative algorithm for solving this dual SVM optimization problem.

Random forest is a tree-based classification algorithm that merges several decision trees to perform the classification task. Each tree is created deterministically using a bootstrap sample of the observation. Additionally, a best split in each node is chosen from a random subset of the predictors rather than all of them. Bagging is a "bootstrap" [10] ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set. Each classifier's training set is generated by randomly drawing, with replacement, N examples, where N is the size of the original training set; many of the original examples may be repeated in the resulting training set while others may be left out. Each individual classifier in the ensemble is generated with a different random sampling of the training set.

In each run, we train the system over 66% of the data and test the system over remaining 33% percent of the data. Testing processes the ads not used for training and decides whether the ads are either scam or trustworthy. In order to compare the results of these classification algorithms we calculated the rank product of F-score, ROC Area, and Accuracy.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)}$$
$$F - Score = \frac{2TP}{(2TP+FP+FN)}$$
$$Precision = \frac{TP}{(TP+FP)}$$

$$Recall = \frac{TP}{(TP+FN)}$$

where TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives.

6.4 Results

We run the classifiers twice, once over the textual features, and then over the textual features plus entities to see what improvement is obtained by using the features we identified in Section 5. For the training phase, we used balanced corpora of scam and trustworthy ads. We choose at random 4244 scam advertisements and 4244 trustworthy advertisements. We report our results for each classifier as follows. In the random forest, the number of trees is 10 and maximum depth is unlimited. With SMO we used the polynomial kernel. For the Bagging algorithm, we get the best results with random forest. Tables 7 and 8 shows these results for bigrams and SPs respectively. As we can, with only textual features, the best accuracy obtained for bigrams is 88.42% for the Bagging classifier. When entities are added, SMO provides an accuracy of 99.06%, with similar performance for all the other measures. With SPs and entities, SMO provides similar performance with 99.2% accuracy.

Classifier		TP	FN	FP	TN	Precision	Recall	Accuracy	F measure	ROC Area
Bagging with Random forest	bigram	1358	112	222	1193	88.6	88.4	88.42	88.4	94.2
	bigram +Entity	1462	8	58	1357	97.8	97.7	97.7	97.7	97.7
SMO	bigram	1308	162	210	1205	87.1	87.1	87.1	87.1	87.1
	bigram + Entity	**1463**	**7**	**20**	**1395**	**99.1**	**99.1**	**99.06**	**99.1**	**99.1**
Naive-Bayes	bigram	1367	103	701	714	76.5	72.1	72.13	70.8	83.2
	bigram +Entity	1380	90	207	1208	90	89.7	89.7	89.7	93.2

Table 7: Results for Bigram features.

Classifier		TP	FN	FP	TN	Precision	Recall	Accuracy	F measure	ROC Area
Bagging with Random forest	SP	1401	69	223	1192	90.3	89.9	89.87	89.8	94.9
	SP +Entity	1462	8	57	1358	97.8	97.7	97.74	97.7	99.6
SMO	**SP + Entity**	**1465**	**5**	**19**	**1396**	**99.2**	**99.2**	**99.16**	**99.2**	**99.2**
Naive-Bayes	SP	1372	98	725	690	76.3	71.5	71.47	69.9	83.9
	SP+Entity	1384	86	206	1209	90	89.9	89.8	89.9	93.8

Table 8: Results for SP features.

7 Related Work

The topic of fraud on the Web is vast. Here, we focus only on fraud as it relates to public electronic market places, such as eBay. We discuss research in fraud detection and stylometric analysis of text as they relate most closely to our work.

Fraud detection techniques. Fraud detection in electronic market places has thus far focused only on auction sites such as eBay. Chau and Faloutsos identify fraudsters on eBay by determining characteristic features from exposed fraudsters and applying a decision tree based classifier to identify other potential fraudsters [5]. The proposed decision tree is trained on 17 features based on prices of items bought or sold in certain time period. This study is consistent with our finding that the offered price by scammers is often lower than the market price. In a follow up work, Chau et al. model the techniques that fraudsters typically use to carry out fraudulent activities, and used data mining and machine learning techniques to spot unnatural patterns in auctions [6]. Their techniques are not applicable to Craigslist scam since there is not auctioning. Subsequently, Pandit et al. [17] looked at trust and authority propagation via user reputation on eBay. Their system, Netprobe, is based on a large ever-changing graph of relationships between eBay users. Based on this graph NetProbe can determine trustworthiness of a particular user. Primarily trustworthiness is determined by how many facilitators(users who boots others' ratings) any given user is associated with. This work sheds light on scam operations but unfortunately can not help in our Craigslist study because Craigslist is anonymous by design.

Almendra and Schwabe [8] studied non-delivery scam on a Brazilian auction site. Their work looks at cell phone fraud and finds that scam is usually characterized by low price and an attractive product.

Fraud campaigns usually were discovered within two weeks of their start but nothing could be done to prevent scam because by the time it was noticed, it was much too late to help the victims. This study is also consistent with our hypothesis that scam is primarily perpetrated by selling an attractive product at an unusually low price. Maranzato et al. apply machine learning for detecting fraudulent behavior of users against reputation systems in the same Brazilian auction site [15, 16]. They detect fraudulent behavior of users against reputation systems by applying logistic regression on real data Such techniques may be applicable to Craigslist spam to detect activity where users mark other people's ads as scam.

Another fraud detection system [13] is designed by visualizing transaction networks on Internet auctions. This system employs link mining techniques and user information to detect suspiciousness. While they have access to user information and the relationships between seller and buyer, they represented the communication through the graph and determined abnormal behavior in the network. However, in Craigslist sellers and buyers communicate offline by email or phone; thus there is no evidence about actual trading between users that can be characterized.

Stylometry. Stylometric techniques attempt to learn an author's writing style to recognize more texts from the same author. We explored whether stylometric techniques could be used to identify writing styles for scam campaigns or learn the writing style of automatic ad posting tools and thus help in scam detection.

Van Halteren [11] focuses on authorship matching using lexicon and syntactic features. Though this system had only 3% error, it relies on a fairly large text to be able to make its attribution. In the Craigslist scenario, we have a few sentences at best and no certain authorship. Brennan and Greenstadt [4] studied a series of three essays written by the same author to find out how easy or difficult it would be to forge an stylometric identity or to make oneself anonymous by intentionally changing one's writing style and find that it is quite easy for a person to change his or her writing style dramatically. In general, stylometric analysis requires a relatively large sample of writing to make attribution and an active attempt to anonymize writing style is likely to be very successful. This suggests that scammers can easily evade detection via stylometric analysis by keeping their ads short and by consciously varying writing style. All ads are already short, making it only necessary for them to alter writing style.

Abbasi [1] developed the Writeprints authorship attribution system, an unsupervised learning method for identification and similarity detection. The objective of this work was to apply author recognition to online documents. The study expands past work by focusing on similarity detection among an unknown set of authors rather than matching a new piece of text to one of previously known authors. Their approach is designed to be scalable across domains such as e-mail, forums, chat. However, it can be defeated by imitating.

While our context was different, we were able to apply lexical features (word or character-based), content-based features (specific words, n-grams) and syntactic features (punctuation) used in above works in our classification process.

8 Conclusions and Future Work

While our study exposed many facets of automobile scam on Craigslist, it opened up many future avenues of exploration. Among the first one is to understand scam in non U.S. markets and in sections other than cars+trucks. Even within automobile scams, many avenues of exploration remain. We highlight them next.

Unidentified campaigns We used various heuristics to find new scam ads belonging to campaigns identified in the scam data set in Section 4. However, this does not find new scam campaigns for which no flagged ads were available. We tested if applying the same criteria as we applied to establish ground truth and to find sister ads of known scam campaigns would help find new campaigns in the `unlabeled` data set but the number of matches were too many to test, owing in part to the large size of that data set. Given this difficulty and the possibility that our scam data set may only contain a fraction of scam on Craigslist, our current investigation does not help answer the question "what percentage of ads on Craigslist are scam"? To answer that question, an exploration of how best to find new scam campaigns is required.

Identifying scammers instead of campaigns We defined scam campaigns as group of ads with same make, model, year and similar price. This definition helps identify similar ads for the same vehicle but not ads that are stylistically similar but may be for different vehicles. The reason that it may be interesting to group ads by style instead of vehicles is because it can help identify scammers behind these ads. There are two approaches that we intend to pursue this goal. The first is to use stylometric techniques and the second is to e-mail scammers and analyze their responses.

Additional features of individual scams While sampling individual campaigns, we found that scammers were using other clever tricks we would like to systematically evaluate. For example, we found the use of images for listing email addresses, foregoing craigslist handles so that craigslist does not have their contact information, and putting additional light or tiny (invisible) text to avoid getting the scam automatically blacklisted.

References

[1] A. Abbasi and H. Chen. A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26, March 2008.

[2] Alexa. `http://www.alexa.com/siteinfo/craigslist.org`.

[3] L. Breiman. Random forests. *Mach. Learn.*, 45:5–32, October 2001.

[4] M. Brennan and R. Greenstadt. Practical attacks against authorship recognition techniques. *Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference*, 2009.

[5] D. H. Chau and C. Faloutsos. Fraud detection in electronic auction. In *In European Web Mining Forum at ECML/PKDD*, 2005.

[6] D. H. Chau, S. P, and C. Faloutsos. Detecting fraudulent personalities in networks of online auctioneers. In *In Proc. ECML/PKDD*, pages 103–114, 2006.

[7] Craigslist scam avoidance tips. `http://www.craigslist.org/about/scams`.

[8] V. da Silva Almendra and D. Schwabe. Analysis of fraudulent activity in a Brazilian auction site. In *International World Wide Web Conference*, April 2024, 2009, Madrid, Spain.

[9] One of the largest eBay scams finally comes to an end. `http://fourpastfour.com/2010/02/24/one-of-the-largest-ebay-scams-finally-comes-to-an-end`, 2010.

[10] B. Efron and R. J. Tibshirani. An introduction to the bootstrap. *Biosystems*, 1971.

[11] H. V. HALTEREN. Author verification by linguistic proling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing*, 4(1), January 2007.

[12] Kelley blue book. `http://www.kbb.com/`.

[13] M. Kobayashi and T. Ito. A transactional relationship visualization system in internet auctions. In *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, IAT '07, 2007.

[14] A. Kolchinsky, A. Abi-Haidar, J. Kaur, A. A. Hamed, and L. M. Rocha. Classification of protein-protein interaction full-text documents using text and citation network features. *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, PP(99):1–1, 05/2010 2010.

[15] R. Maranzato, M. Neubert, A. M. Pereira, and A. P. do Lago. Feature extraction for fraud detection in electronic marketplaces. In *LA-WEB/CLIHC*, 2009.

[16] R. Maranzato, A. M. Pereira, A. P. do Lago, and M. Neubert. Fraud detection in reputation systems in e-markets using logistic regression. In *SAC*, 2010.

[17] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos. NetProbe: A fast and scalable system for fraud detection in online auction networks. In *International World Wide Web Conference*, 2007, May 8–12, 2007, Banff, Alberta, Canada.

[18] J. C. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[19] K. Sparck Jones and P. Willett, editors. *Readings in information retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[20] A. Yoskowitz. Craigslist is 'cesspool of crime,' says rival. `http://www.afterdawn.com/news/article.cfm/2011/02/26/craigslist_is_cesspool_of_crime_says_rival`, 2011.

A Additional Tables and Figures

unigram	featur, system, titl, contact, pleas, conveni, control, 2008, need, vehicl, truck, rear, honda, nissan, chevi, nice, plenti, seat, heat
sp	conveni–featur, with–conveni,automat–with,clean–titl, featur–power,clean–leather,nice–clean,plenti–rubber, rubber–left,tire–plenti, left–alloi, nice–also, wheel–nice,increas–test,alloi–wheel,drive–automat, leather–interior,test–drive,steer–wheel,fulli–paid
bigram	conveni–featur, with–conveni,automat–with, featur–power, clean–titl, clean–leather, nice–clean, rubber–left, plenti–rubber, tire–plenti, left–alloi, nice–also, wheel–nice, increas–test, alloi–wheel, drive–automat, leather–interior, test–drive, steer–wheel,fulli–paid

Table 9: The top 20 unigram, sp and bigram features

2007 dodge durango sxt 4x4 flex ethenol and gasoline - $13800 (woodland hills)

Date: 2011-04-01, 5:03PM PDT
Reply to: sale-gz3m7-2300183753@craigslist.org [Errors when replying to ads?]

2007 dodge durango sxt 4x4 super power flex fluid you can use gasoline or Ethanol ,super economic , is ready paid ,registration paid until 2012 may ,clean title , transmission automatic, power steering,power door locks,key entry,am/FM stereo ,CD,tinted glass ,rear defroster, seats material cloth ,power brakes,cruise control,air condition,excellent condition,one owner,never accident ,looks new ,call 1818 9176060 miles 74,0000

- Location: woodland hills
- it's NOT ok to contact this poster with services or other commercial interests

PostingID: 2300183753

For Sale 2007 Dodge Durango 4.7 V8 - $11000 (Granbury)

Date: 2011-04-18, 8:08AM CDT
Reply to: sale-cnsky-2331887419@craigslist.org [Errors when replying to ads?]

For Sale 07 Dodge Durango 4.7 V8, power windows door locks, cd player front an rear air controls, 3rd row seat, 18" wheels, tinted windows, after market flowmaster exhaust, 43,994 miles runs great, need to sale asap $11,000.00 OBO, 817-894-1677

- Location: Granbury
- it's NOT ok to contact this poster with services or other commercial interests

PostingID: 2331887419

Figure 12: Two good ads from the deleted data set, incorrect found as potential sister ads.

	Feature	probP	probN	score
bigram	conveni–featur	0.00024	0.12276	1
	with–conveni	0.00000	0.12088	2
	automat–with	0.01084	0.12677	3
	featur–power	0.00094	0.10391	4
	clean–titl	0.14828	0.04618	5
	clean–leather	0.00306	0.10203	6
	nice–clean	0.00990	0.10462	7
	rubber–left	0.00000	0.09025	8
	plenti–rubber	0.00000	0.09025	9
	tire–plenti	0.00047	0.09048	10
sp	conveni–featur	0.00024	0.12276	1
	with–conveni	0.00000	0.12088	2
	automat–with	0.01273	0.12771	3
	clean–titl	0.14946	0.04642	4
	featur–power	0.00141	0.10391	5
	clean–leather	0.00566	0.10203	6
	nice–clean	0.01061	0.10485	7
	plenti–rubber	0.00000	0.09025	8
	rubber–left	0.00000	0.09025	9
	tire–plenti	0.00047	0.09048	10

Table 10: The positive/negative probability and s score of top 10 sp and bigram features.