Crawford et al. Journal of Big Data (2015) 2:23
DOI 10.1186/s40537-015-0029-9

Journal of Big Data
a SpringerOpen Journal

SURVEY PAPER Open Access

Survey of review spam detection using machine learning techniques

Michael Crawford*, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter and Hamzah Al Najada

* Correspondence:
michaelcrawf2014@fau.edu
Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA

Abstract

Online reviews are often the primary factor in a customer's decision to purchase a product or service, and are a valuable source of information that can be used to determine public opinion on these products or services. Because of their impact, manufacturers and retailers are highly concerned with customer feedback and reviews. Reliance on online reviews gives rise to the potential concern that wrongdoers may create false reviews to artificially promote or devalue products and services. This practice is known as Opinion (Review) Spam, where spammers manipulate and poison reviews (i.e., making fake, untruthful, or deceptive reviews) for profit or gain. Since not all online reviews are truthful and trustworthy, it is important to develop techniques for detecting review spam. By extracting meaningful features from the text using Natural Language Processing (NLP), it is possible to conduct review spam detection using various machine learning techniques. Additionally, reviewer information, apart from the text itself, can be used to aid in this process. In this paper, we survey the prominent machine learning techniques that have been proposed to solve the problem of review spam detection and the performance of different approaches for classification and detection of review spam. The majority of current research has focused on supervised learning methods, which require labeled data, a scarcity when it comes to online review spam. Research on methods for Big Data are of interest, since there are millions of online reviews, with many more being generated daily. To date, we have not found any papers that study the effects of Big Data analytics for review spam detection. The primary goal of this paper is to provide a strong and comprehensive comparative study of current research on detecting review spam using various machine learning techniques and to devise methodology for conducting further investigation.

Keywords: Review spam; Opinion mining; Web mining; Machine learning; Big data; Classification

Introduction

As the Internet continues to grow in both size and importance, the quantity and impact of online reviews continually increases. Reviews can influence people across a broad spectrum of industries, but are particularly important in the realm of e-commerce, where comments and reviews regarding products and services are often the most convenient, if not the only, way for a buyer to make a decision on whether or not to buy them. Online reviews may be generated for a variety of reasons. Often, in an effort to improve and enhance their businesses, online retailers and service

providers may ask their customers to provide feedback about their experience with the products or services they have bought, and whether they were satisfied or not. Customers may also feel inclined to review a product or service if they had an exceptionally good or bad experience with it. While online reviews can be helpful, blind trust of these reviews is dangerous for both the seller and buyer. Many look at online reviews before placing any online order; however, the reviews may be poisoned or faked for profit or gain, thus any decision based on online reviews must be made cautiously. Furthermore, business owners might give incentives to whoever writes good reviews about their merchandise, or might pay someone to write bad reviews about their competitor's products or services. These fake reviews are considered review spam and can have a great impact in the online marketplace due to the importance of reviews.

Review spam can also negatively impact businesses due to loss in consumer trust. The issue is severe enough to have attracted the attention of mainstream media and governments. For example, the BBC and New York Times have reported that "fake reviews are becoming a common problem on the Web, and a photography company was recently subjected to hundreds of defamatory consumer reviews" [1]. In 2014, the Canadian Government issued a warning "encouraging consumers to be wary of fake online endorsements that give the impression that they have been made by ordinary consumers" and estimated that a third of all online reviews were fake[1]. As review spam is a pervasive and damaging problem, developing methods to help businesses and consumers distinguish truthful reviews from fake ones is an important, but challenging problem.

In the literature, review spam has been categorized into three groups, proposed by Dixit et al. [2]: (1) Untruthful Reviews – the main concern of this paper, (2) Reviews on Brands – where the comments are only concerned with the brand or the seller of the product and fail to review the product, and (3) Non-Reviews – those reviews that contain either unrelated text or advertisements. The first category, untruthful reviews, is of most concern as they undermine the integrity of the online review system. Detection of type 1 review spam is a challenging task as it is difficult, if not impossible, to distinguish between fake and real reviews by manually reading them. To illustrate the difficulty of this task, we consider a real and fake example from the dataset created by Ott et al. [3]. As a human judge it is difficult to confidently ascertain which review is fake and which is authentic.

Review 1: *Great Hotel This building has been fantastically converted into studios/suites. We only had a studio which was brilliant can't imagine how the suite could have bettered what we had. The kitchen had everything cooker microwave dishwasher and fridge freezer. Bathroom was a good size and again had everything you need including good quality toiletries. Hotel also has a good gym and swimming pool and excellent laundry facilities if you need them. The complimentary breakfast each morning was also very good and had an excellent choice. The parking in the hotel was secure and reasonably priced. The location was pretty central and had easy access to the underground city. Would definitely stay here again.*

Review 2: *During my latest business trip, both me and my wife recently stayed at the Omni Chicago Hotel in Chicago, Illinois, at one of their Deluxe suites. Unfortunately, and I think I speak for both of us, we were not fully satisfied with the hotel. The hotel advertises luxury-level accommodations, and while the rooms resemble what one can*

see in the pictures, the service is certainly sub-par. When one plans a stay at such an establishment, they expect a service that goes beyond having fresh towels in the bathroom when they check in. First of all, the air-conditioning in the room seemed to be in need of a new filter and when it was first turned on, the air coming out seemed musty. Second of all, the fitness center was only open until 10:30 pm. For people who like to exercise after dinner, this can certainly be a problem. Especiaally considering that it does not take much to have the fitness center available around the clock or until midnight. For these, as well as other similar reasons, I would not recommend this hotel, if one is looking for luxury accommodations.

There are no clear indications or signals from the text of the two reviews that indicate to the casual reader that the first review is real while the second is a fake. Nevertheless, guides provided by the Consumerist[2] and MoneyTalksNews[3] websites offer tips to help consumers spot fake reviews. A computer scientist might seek to utilize this logic when training data mining and machine learning algorithms to find these features in the review that will determine if it is real or fake.

Over 18 million reviews were created on Yelp 2014[4] and Trip Advisor currently has over 200 million reviews[5]. Online reviews are constantly being generated on various web sites across the Internet. Consequently, Big Data techniques are needed to address the problem of review spam. Big Data, while an overused buzzword with an elusive definition, is often quantified with the Four V's[6]: (1) Volume – the sheer size and scale of the data, (2) Velocity – the rate at which new data is created and consumed by processing engines, (3) Variety – the different formats that data may be stored in, and (4) Veracity – the quality level of the data. The Volume and Velocity of online reviews are noted by merely visiting e-commerce and customer rating sites, such as Yelp and Amazon. There is great Variety across the possible industry sectors for reviews (such as hotels, restaurants, e-commerce, home services, etc.), along with the multiplicity of languages that reviews are written in. Veracity is a problem with online reviews, since the vast majority of reviews are unlabeled, which means it is not easily known whether the review is fake or not. Additionally, standard machine learning algorithms tend to break down and become ineffective when dealing with data of this size, which poses a problem when trying to apply these algorithms for review spam detection [4]. Thus, review spam detection is a Big Data problem, as there are numerous challenges when analyzing and classifying varying reviews from disconnected sources.

Data mining and machine learning techniques, primarily those for web and text mining, offer an exciting contribution to detecting fraudulent reviews. According to Liu [5], web mining is "the process for finding useful information and relations from the contents available on the web by largely relying on the available machine learning techniques and methods". Web mining can be divided into three types of tasks: structure, content and usage mining. Content mining is concerned with knowledge and information extraction, and categorizing entities using data mining and machine learning approaches. A straightforward example of content mining is opinion mining. Opinion mining consists of attempting to ascertain the sentiment (i.e., positive or negative polarity) of a text passage by analyzing the features of that passage. A classifier can be trained to classify new instances by analyzing the text features associated with different opinions along with their sentiment. Review spam detection, like opinion mining, lies in the category of content mining, but also utilizes features not directly linked to the

content [6]. Constructing features to describe the text of the review involves text mining and Natural Language Processing (NLP). Additionally, there may be features associated with the review's writer, its post date/time and how the review deviates from other reviews for the same product or service.

It is important to mention that while most existing machine learning techniques are not sufficiently effective for review spam detection, they have been found to be more reliable than manual detection. The primary issue, as identified by Abbasi et al. [7], is the lack of any distinguishing words (features) that can give a definitive clue for classification of reviews as real or fake. A common approach in text mining is to use a bag of words approach where the presence of individual words, or small groups of words are used as features; however, several studies have found that this approach is not sufficient to train a classifier with adequate performance in review spam detection. Therefore, additional methods of feature engineering (extraction) must be explored in an effort to extract a more informative feature set that will improve review spam detection. In the literature, there are many studies that consider different sets of features for the study of review spam detection utilizing a variety of machine learning techniques. Jindal et al. [8], Li et al. [9] and Mukherjee et al. [10], used individual words from the review text as the features, while Shojaee et al. [11] used syntactic and lexical features. An additional study by Ott et al. [12] used review characteristic features in addition to unigram and bigram term-frequencies.

Features associated with the behavior of the reviewer also merit further investigation. The study of writers of review spam differs from that of the review spam itself since features representing the characteristics and behaviors for reviewers cannot be extracted from the text of a single review. Examples of studying spammer behavior include spotting multiple User IDs for the same author [13] and identifying groups of spammers by studying their behavioral footprints [14–16]. Alternatively, graph-theory based methods can also be used to find relationships between the reviews and their corresponding authors and have shown promising results [17, 18]. Combining review spam detection through a review's features, and spammer detection through analysis of their behavior may be a more effective approach for detecting review spam than either approach alone.

Before addressing the challenges associated with improving review spam detection, we must first address collection of data. Data is a major part of any machine learning based model, and while a massive volume of reviews are available on the Internet, collecting and labeling a sufficient number of them to train a review spam classifier is a difficult task. An alternative to collecting and labeling data is to artificially create review spam datasets by using synthetic review spamming, which takes existing truthful reviews and builds fake reviews from them. Sun et al. [19] used this approach to create a review spam dataset.

In this paper we discuss machine learning techniques that have been proposed for the detection of online review spam, with an emphasis on feature engineering and the impact of those features on the performance of the spam detectors. Additionally, the merits of supervised, unsupervised and semi-supervised learning methods are analyzed and results of current research using each approach presented along with a comparative analysis. Finally, we provide suggestions for aspects of review spam detection requiring further investigation, and best practices for conducting future research. To

the best of our knowledge, this paper includes information about all of the datasets that have been used, or generated for use, in the reviewed literature.

The structure of this paper is as follows. The Feature Engineering for Review Spam Detection section provides an overview of feature engineering in this domain, both for review centric spam detection and reviewer centric spam detection. The Review Centric Review Spam Detection section discusses and analyzes current research using supervised, unsupervised and semi-supervised machine learning for review centric spam detection. The Reviewer Centric Review Spam Detection section provides an overview of studies using reviewer centric features. The Comparative Analysis and Suggestions section contains a discussion and comparison between the different methods proposed. The Conclusion summarizes our findings and reviews the important of both past and future work.

Feature engineering for review spam detection

Feature engineering is the construction or extraction of features from data. In this section, we analyze and discuss some of the commonly used features in the domain of review spam detection. As briefly outlined in the introduction, previous studies have used several different types of features that can be extracted from reviews, the most common being words found in the review's text. This is commonly implemented using the bag of words approach, where features for each review consist of either individual words or small groups of words found in the review's text. Less frequently, researchers have used other characteristics of the reviews, reviewers and products, such as syntactical and lexical features [11] or features describing reviewer behavior. The features can be broken down into the two categories of review and reviewer centric features. Review centric features are features that are constructed using the information contained in a single review. Conversely, reviewer centric features take a holistic look at all of the reviews written by any particular author, along with information about the particular author.

It is possible to use multiple types of features from within a given category, such as bag-of-words with POS tags, or even create feature sets that take features from both the review centric and reviewer centric categories. Using an amalgam of features to train a classifier has generally yielded better performance then any single type of feature, as demonstrated in Jindal et al. [20], Jindal et al. [21], Li et al. [9], Fei. et al. [22], Mukherjee et al. [23] and Hammad [24]. Li et al. [25] concluded that using more general features (e.g., LIWC and POS) in combination with bag-of-words, is a more robust approach than bag-of-words alone. A study by Mukherjee et al. [23] found that using the abnormal behavioral features of the reviewers performed better than the linguistic features of the reviews themselves. The following subsections discuss and provide examples of some review centric and reviewer centric features.

Review centric features

We split review centric features into several categories. First, we have bag-of-words, and bag-of-words combined with term frequency features. Next, we have Linguistic Inquiry and Word Count (LIWC) output, parts of speech (POS) tag frequencies,

Stylometric and Syntactic features. Finally, we have review characteristic features that refer to information about the review not extracted from the text.

Bag of words

In a bag of words approach, individual or small groups of words from the text are used as features. These features are called n-grams and are made by selecting n contiguous words from a given sequence, i.e., selecting one, two or three contiguous words from a text. These are denoted as a unigram, bigram, and trigram (n = 1, 2 and 3) respectively. These features are used by Jindal et al. [21], Li et al. [9] and Fei et al. [22]. However, Fei et al. observed that using n-gram features alone proved inadequate for supervised learning when learners were trained using synthetic fake reviews, since the features being created were not present in real-world fake reviews. An example of the unigram text features extracted from three sample reviews is shown in Table 1. Each occurrence of a word within a review will be represented by a "1" if it exists in that review and "0" otherwise.

1. ***Review1****: The hotel rooms were so great*
2. ***Review2****: We had a great time at this hotel great stay*
3. ***Review3****: The rooms service is bad*

Term frequency

These features are similar to bag of words but also include term-frequencies. They have been used by Ott et al. [12] and Jindal et al. [8]. The structure of a dataset that uses the term frequencies is shown in Table 2, and is similar to that of the bag of words dataset; however, instead of simply being concerned with the presence or absence of a term, we are concerned with the frequency with which a term occurs in each review, so we include the count of occurrences of a term in the review.

4. ***Review4****: The hotel rooms were so great, were so comfort*
5. ***Review5****: We had a great time at this hotel great stay*
6. ***Review6****: The rooms service is bad so bad*

LIWC output and POS tag frequencies

Linguistic Inquiry and Word Count[7] (LIWC) is a text analysis software tool in which users can "build [their] own dictionaries to analyze dimensions of language specifically relevant to [their] interests." Part of Speech (POS) tagging involves tagging word features with a part of speech based on the definition and its context within the sentence in which it is found [26]. Ott et al. [3] and Li et al. [25] achieved better results by also including these features than with bag of words alone. Table 3 shows the results from

Table 1 Example of text features dataset structure, for reviews 1, 2 and 3

Review	the	hotel	rooms	were	so	great	we	had	a	time	at	this	service	is	bad	stay
Review1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Review2	0	1	0	0	0	1	1	1	1	1	1	1	0	0	0	1
Review3	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0

Table 2 Example of text features frequencies dataset structure, for reviews 4, 5 and 6

Review	the	hotel	rooms	were	so	great	comfort	we	had	a	time	at	this	service	is	bad	stay
Review4	1	1	1	2	2	1	1	0	0	0	0	0	0	0	0	0	0
Review5	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0	0	1
Review6	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	2	0

the LIWC program when applied to Review 7. Personal text refers to text associated with personal concerns such as work, home or leisure activities. Formal text refers to text disassociated from personal concerns, consisting of psychological processes, linguistic processes and spoken categories. Below Review 7 is the review along with POS tags for each word. Table 4 shows the meaning of each POS tag[8], while Table 5 presents the frequencies of these tags within the review.

7. ***Review7****: I like the hotel so much, the hotel rooms were so great, the room service was prompt, I will go back for this hotel next year. I love it so much. I recommend this hotel for all of my friends.*
   Review7: I_PRP like_VBP the_DT hotel_NN so_RB much_RB,_, The_DT hotel_NN rooms_NNS were_VBD so_RB great_JJ,_, the_DT room_NN service_NN was_VBD prompt_JJ,_, I_PRP will_MD go_VB back_RB for_IN this_DT hotel_NN next_JJ year_NN ._. I_PRP love_VBP it_PRP so_RB much_RB ._. I_PRP recommend_VBP this_DT hotel_NN for_IN all_DT of_IN my_PRP$ friends_NNS ._.

Stylometric

These features were used by Shojaee et al. [11] and are either character and word-based lexical features or syntactic features. Lexical features give an indication of the types of words and characters that the writer likes to use and includes features such as number of upper case characters or average word length. Syntactic features try to "represent the writing style of the reviewer" and include features like the amount of punctuation or number of function words such as "a", "the", and "of".

Semantic

These features deal with the underlying meaning or concepts of the words and are used by Raymond et al. [1] to create semantic language models for detecting

Table 3 LIWC results when applying Review7 text

LIWC Dimension	Your data	Personal texts	Formal texts
Self-references (I, me, my)	12.50	11.4	4.2
Social words (Mate, talk, they, child)	2.50	9.5	8.0
Positive emotions (Love, nice, sweet)	5.00	2.7	2.6
Negative emotions (Hurt, ugly, nasty)	0.00	2.6	1.6
Overall cognitive words (cause, know, ought)	0.00	7.8	5.4
Articles (a, an, the)	7.50	5.0	7.2
Big words (>6 letters)	7.50	13.1	19.6

Table 4 POS tags abbreviation descriptions

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

untruthful reviews. The rationale is that changing a word like “love” to “like” in a review should not affect the similarity of the reviews since they have similar meanings.

Review characteristic

These features contain metadata (information about the reviews) rather than information on the text content of the review and are seen in works by Li et al. [9] and Hammad [24]. These characteristics could be the review’s length, date, time, rating, reviewer id, review id, store id or feedback. An example of review characteristic features is presented in Table 6. Review characteristic features have shown to be beneficial in review spam detection. Strange or anomalous reviews can be identified using this metadata, and once a reviewer has been identified as writing spam it is easy to label all reviews associated with their reviewer ID as spam. Some of these features may not be available for all sources of review spam and thus limits their utility for detection of spam in many data sources.

Reviewer centric features

As highlighted earlier, identifying spammers can improve detection of fake reviews, since many spammers share profile characteristics and activity patterns. Various combinations of features engineered from reviewer profile characteristics and behavioral patterns have been studied, including work by Jindal et al. [20], Jindal et al. [21], Li et al. [9], Fei et al. [22], Mayzlin et al. [27] and Mukherjee et al. [23]. Examples of reviewer centric features are presented in Table 7 and further elaboration on select features used in Mukherjee et al. [23] along with some of their observations follows:

Table 5 POS tagging frequencies for Review 7

POS Tag		DT	IN	JJ	MD	NN	NNS	PRP	RB	VB	VBD	VBP
Review	Review 7	6	3	3	1	7	2	6	6	1	2	3

Maximum number of reviews

It was observed that about 75 % of spammers write more than 5 reviews on any given day. Therefore, taking into account the number of reviews a user writes per day can help detect spammers since 90 % of legitimate reviewers never create more than one review on any given day.

Percentage of positive reviews

Approximately 85 % of spammers wrote more than 80 % of their reviews as positive reviews, thus a high percentage of positive reviews might be an indication of an untrustworthy reviewer.

Review length

The average review length may be an important indication of reviewers with questionable intentions since about 80 % of spammers have no reviews longer than 135 words while more than 92 % of reliable reviewers have an average review length of greater than 200 words.

Reviewer deviation

It was observed that spammers' ratings tend to deviate from the average review rating at a far higher rate than legitimate reviewers, thus identifying user rating deviations may help in detection of dishonest reviewers.

Maximum content similarity

The presence of similar reviews for different products by the same reviewer has been shown to be a strong indication of a spammer. Mukherjee et al. [23] used cosine similarity; however, other more advanced similarity functions based upon word meanings versus the words themselves have shown promise [1].

Review centric review spam detection

Review centric review spam detection is the most common form of review spam detection, which uses machine learning techniques to build models using the content and metadata of the reviews. Supervised learning refers to the task of learning from

Table 6 Reviews characteristics dataset structure

Review	Review ID	Product ID	Reviewer ID	Rating	Helpfulness	Review char	Review words	Date	Time
Review4	152	012345	226	1	1	38	9	8/5/2013	09:24
Review5	153	012345	789	5	0	35	10	9/1/2015	12:06
Review6	154	012345	789	5	0	25	7	9/1/2015	12:07

Table 7 Reviewers characteristics dataset structure

Reviewer#	Product ID	Reviewer ID	Reviewer name	Email address	# of Reviews	First review	Last review	Max # reviews per day	Average rating	Date	Time
Reviewer1	123456	152	JO	jo@gmail	2000	09/01/13	09/30/14	30	5	09/30/14	12:05
Reviewer2	123456	153	LI	jo@gmail	2300	09/01/13	09/30/14	31	5	09/30/14	12:06
Reviewer3	123456	154	SA	sa@gmail	3	05/02/11	06/05/14	1	4	06/05/14	12:00

labeled data and is the most prevalent method used for review spam detection in the literature. Unfortunately, this method requires labeled data in order to train a classifier, presenting the challenge of needing methods to procure and accurately label a sufficient amount of data, which can be problematic in the field of review spam detection. Conversely, unsupervised learning uses unlabeled data to find unseen relationships between instances independent of a class attribute. An example of unsupervised learning is clustering, which is able to group instances of unlabeled data based upon some type of similarity function. Semi-supervised learning is a combination of the two and uses a few labeled instances in combination with a large number of unlabeled instances to train a classifier and has shown promise in the area of review spam detection. These methods are summarized in Table 8 and the following subsections outline research conducted using these different types of learning in the domain of review spam detection.

Supervised learning

Supervised learning can be used to detect review spam by looking at it as the classification problem of separating reviews into two classes: spam and non-spam reviews. To the best of our knowledge, the first researchers to have studied deceptive opinion spam using supervised learning were Jindal et al. [21]. They discuss the evolution of opinion mining, which had primarily focused on extracting or summarizing the opinions from text by using Natural Language Processing (NLP). Prior to their contribution, the content characteristics of the text that might indicate abnormal activities, such as creating review spam, had not been addressed. In an effort to investigate opinion spam in reviews and devise techniques for review spam detection, Jindal et al. collected 5.8

Table 8 Types of machine learning techniques

Method	Attributes
Supervised Learning	Learning from a set of labeled data
	Requires labeled training data
	Most common form of learning
Unsupervised Learning	Learning from a set of unlabeled data
	Finds unseen relationships in the data independent of class label
	Most common form is clustering
Semi-supervised Learning	Learning from labeled and unlabeled data
	Only requires a relatively small set of labeled data which is supplemented with a large amount of unlabeled data
	Ideal for cases such as review spam where vast amounts of unlabeled data exist

million reviews of products on Amazon generated by 2.14 million users. The authors categorized the reviews of class spam into three types: untruthful opinion, reviews on brand only, and non-review (labeled types 1, 2 and 3 respectively). They started by finding the near duplicate reviews, which they defined as reviews with a Jaccard similarity score of over 90 % of their 2-g. This was done using a method known as w-shingling[9]. An alternate method for detecting near duplicates using Symantec Language Models (SLM) was developed by Raymond et al. [1]. They then extracted 36 additional features that describe the review, reviewer and product reviewed. A logistic regression model was built using these examples, and when tested using 10-fold cross validation, an Area Under the receiver operating characteristic Curve (AUC) score of 0.78 was achieved when using all features, compared to an AUC score of 0.63 when only using text features. However, the authors recognized that simply finding duplicate reviews is a trivial task and they wanted to test if a model trained using duplicate reviews would generalize to find review spam in general. They manually analyzed 100 non-duplicate reviews that the classifier predicted were spam with the most confidence and found that 52 % were clearly spam. It was also hypothesized that outlier reviews may in fact also be opinion spam and a series of lift curves was constructed to demonstrate the classifier's effectiveness in identifying these outliers as potential spam.

As a further test, to compare these results with detection of type 2 and 3 review spam, they manually labeled 470 instances of these types of spam and trained a logistic regression classifier using them. They also tried to use Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers but found they did not perform as well. The best model, evaluated using 10-fold cross validation, achieved an AUC score of over 98 %. From this, they concluded that review spam of types 2 and 3 are much easier to spot and focus should be paid on type 1 (untruthful opinion) review spam. Using text only features resulted in an AUC score of 90 % for detection of type 2 and 3 review spam. Their work shows that text features alone are insufficient for detection of review spam, and the addition of other types of features often improves results; however, as more types of features are extracted it can be expected that feature set size increases along with the training dataset size, making the training of a classifier more computationally expensive and also possibly leading to over fitting. Further work should also investigate feature selection techniques as a means of reducing data dimensionality and improving classifier performance. Feature selection selects an optimal subset of features, removing redundant and irrelevant features that may be detrimental to classification performance, or result in over-fitting [28]. Additionally, by reducing the number of features used to train a model, the computational complexity of the task is reduced.

Ott et al. [3] developed and compared three approaches for performing deceptive spam detection. For their study, they produced a new dataset using Amazon Mechanical Turk (AMT)[10] in combination with TripAdvisor. The untruthful reviews were created by requesting a group of people to deliberately write 400 fake reviews of positive sentiment (i.e., 5 star reviews) for a set of hotels. Additionally, 400 "truthful" 5-star reviews were collected from the TripAdvisor website for the same hotels. The resulting dataset consisted of 800 reviews with positive sentiment towards the hotels (fully balanced with 400 deceptive and 400 truthful-reviews). In a later work they created a second dataset of the same size and similarly balanced, but of negative sentiment (i.e., 1 and

2 star reviews) [12]. Combined together, they claimed this to be the first known "gold-standard" dataset for review spam. For this work, three groups of features were identified: POS tag frequencies, LIWC output [29], and text categorization based features using bigrams. Naïve Bayes and SVM classifiers were trained and evaluated using 5-fold nested cross-validation where all reviews for any given hotel are fully contained within a given fold. Their best model achieved an accuracy of 89.8 % using bigram and LIWC features with an SVM classifier. They also had three human judges evaluate one fold (160 reviews); the highest accuracy score achieved by a human judge was approximately 61 %, showing the classifier to outperform human judges by a substantial margin. It should be noted that while the data set developed in this study is one of the most used datasets for research in review spam detection, it is not necessarily an accurate representation of actual review spam since the fake reviews generated for this data set were written with the intent of being used for research and outsourced to unknown parties, rather than consisting of authentic spam reviews designed with the intent of influencing consumers. It is unclear if a model trained using this dataset will yield similar results when evaluated on real world data.

Ott et al. [12] conducted a more recent study of deceptive opinion spam using the same data and framework as they used earlier [3]; however, they limited their scope to n-gram based features and only used the SVM classifier since it outperformed Naïve Bayes in their earlier work. Using unigram and bigram term frequency features achieved an accuracy of approximately 86 % when considering only reviews with negative sentiment. Again they had human judges evaluate the reviews and found the classifier outperformed them, with the best judge achieving an accuracy of 65 %. They also tested classifier performance when using both the positive and negative reviews together for training and observed that the accuracy on reviews with positive sentiment dropped from 89.3 to 88.4 %. The most notable observation is that doubling the size of the training data, by adding negative sentiment training instances, did not improve results and, in fact, slightly lowered the accuracy of detecting positive sentiment spam reviews. This suggests that separating spam review detection into positive sentiment spam review detection and negative sentiment spam review detection is beneficial. Again, some of the data being used is not real-world data and it remains unknown if the performance of classifiers trained using their data will carry over to purely real world datasets. Additionally, it should be noted that their experiment relied entirely on n-gram features, which were shown by Jindal et al. [21] to be inferior to n-grams in combination with other types of features.

An alternative classification framework was proposed by Li et al. [25]. In their work they argued that existing supervised learning algorithms in literature are usually narrowed to one specific domain and rely heavily on domain-specific vocabulary. To address this, they tried to improve our understanding of the nature of deceptive reviews by creating a cross domain dataset that included three types of reviews from three domains (hotel, restaurant and doctor). AMT was used to solicit fake reviews; however an additional set of fake reviews was solicited from "domain experts". Truthful reviews were collected from review websites. Their classification framework was based on using the Sparse Additive Generative Model (SAGE), which is a generative Bayesian approach introduced by Eisenstein et al. [30]. Basically, it is "a Bayesian generative approach that can capture the multiple generative facets (i.e., deceptive vs. truthful, positive vs.

negative, experienced vs. non-experienced, hotel vs. restaurant vs. doctor)" [25]. The authors used a combination of topic models (statistical models for discovering abstract topics in a collection of documents) and generalized additive models (linear models in which the linear predictor is dependent on unknown, smoother functions) generated using SAGE as well as SVM in their classification experiments. Additionally, they investigated different methods of feature engineering and found the use of more general features, such as LIWC and POS, to be more robust than unigram features alone when modeled using SAGE for cross-domain classification; however, when comparing the intra-domain classification (i.e., hotels reviews only) the best performance is achieved by unigram features. This indicates that different linguistic features may appear in different domains, and more robust cues of deceptive opinion spam need to be identified if a cross domain classifier is to be created. Of note was that the classifier exhibited particular difficulty when trained using the restaurant and hotel reviews and evaluated against the doctor reviews. Using SAGE, accuracies of 64.7 and 63.4 % were achieved using LIWC and POS tag features respectively but only 52.0 % when using Unigram features.

Shojaee et al. [11] proposed a novel method for detecting review spam by using Stylometric (Lexical and Syntactic) features. (For further details on Stylometric features see Abbasi et al. [31]). The features in this work are categorized as either lexical features or syntactic features. Lexical features are character/word based features, while syntactic features represent the writing style of the reviewers at the sentence level, such as occurrences of punctuations or function words. In this work they built SVM and Naïve Bayes classifiers on the dataset created by Ott et al. [3] using a hybrid set of both the lexical and syntactic features and compared this with using either lexical or syntactic features alone. Using 10-fold cross validation, they observed that the hybrid feature set using the SVM learner achieved the highest performance, an F-measure of 84 %. Additionally, SVM outperformed Naïve Bayes for all sets of features. A potential concern of this study is that the model was trained and evaluated on synthetic fake reviews. Due to this, it is possible that the classifier performance measured is a poor indication of real world performance, as was demonstrated by Mukherjee et al. [32]. Also there is no comparison evaluation to determine if using these Stylometric features in addition to n-gram features enhances classification performance.

Review spam can be found in multiple languages, as reviewers from all around the world can write online reviews in any language they want. While many of the features will remain unchanged (i.e., spammers characteristics and behaviors), word features will change to reflect each language. One study by Hammad [24] proposes an approach for spam detection in Arabic opinion reviews, illustrating that the methods used in the above papers can be extended to multiple languages. Hammad, in addition, recognized the imbalanced class distribution of reviews gathered online. An imbalanced class distribution occurs when the class of interest, in this case spam reviews, has relatively few instances compared with the class that is not of interest (non-spam) reviews. Class imbalance makes it more difficult to identify spam reviews as classifiers may be biased towards the majority class. Moreover, Hammad created a new datatset by crawling Arabic reviews from *tripadvisor.com, booking.com, and agoda.ae.* He then manually labeled the data by following a set of rules such as: duplicate and near duplicate reviews are labeled as spam, reviews about brands only are considered as spam, and non-reviews such as ads, discussions, or irrelevant reviews

are labeled as spam. He extracted 26 features for use in his experiments by combining review content (text) features, reviewer features, and hotel information features. As the data he collected was extremely imbalanced, he applied Random Undersampling (RUS) and Random Oversampling (ROS) to alleviate problems associated with data skew. Contrary to other research efforts, he found that Naïve Bayes yielded the best performance, and outperformed SVM. Using ROS he was able to achieve an F-measure score of 99.59 %. This study is important since it was the first to try to address the class imbalance inherent to real world spam detection, in contrast to previous studies that constructed balanced, or roughly balanced data sets.

Unsupervised learning

Because of the difficulty of producing accurately labeled datasets of review spam, the use of supervised learning is not always applicable. Unsupervised learning provides a solution for this, as it doesn't require labeled data. A novel unsupervised text mining model was developed and integrated into a semantic language model for detecting untruthful reviews by Raymond et al. [1] and compared against supervised learning methods. Their model creates an approximation method for calculating the degree of untruthfulness for reviews based on the duplicate identification results by estimating the overlap of semantic contents among reviews using a Semantic Language Model (SLM). In addition to performing unsupervised review spam detection, they also developed a high-order concept of association mining to extract context-sensitive concept association knowledge. Their model follows the assumed logic that if the semantic content of a review is close to those of another review, it is likely that the two reviews are duplicates and thus examples of spam reviews. For their experiment, they built a dataset from real-world reviews collected from Amazon. They first identified reviews with a cosine similarity above some threshold and manually reviewed them to determine if they were indeed spam. Pairs of reviews which were determined to be spam by at least 2 out of 3 human judges were labeled as such, and the rest thrown out. Conversely, reviews that did not have a cosine similarity above a certain threshold with any other reviews were kept as instances of truthful reviews and not manually reviewed. The final dataset contained 54,618 reviews, of which 6 % were spam. Their SLM was then used to assign a "spamminess" score to each instance. Using this score, they were able to achieve an AUC of .9987 while an SVM model trained on the same data achieved an AUC of 0.5571. They argue that their experimental results show that a semantic language modeling and a text mining-based computational model are effective for the detection of untruthful reviews, and that unsupervised methods can achieve a high detection rate of duplicate spam reviews.

It should be noted that the high results achieved by SLM are to be expected, as SLM is performing an operation similar to how the data was labeled. Data was labeled as spam if it had a high cosine similarity with another instance while their model measures similarity between instances using semantic analysis. Further work is needed to address how SLM, and other unsupervised methods, perform on datasets that have review spam which is not similar to other instances and instances of truthful reviews that are similar to other truthful reviews, since their particular dataset gathered spam and non-spam reviews from the two ends of the similarity spectrum, ignoring everything in the middle.

Semi-supervised learning

In other domains, it has been found that using unlabeled data in conjunction with a small amount of labeled data can considerably improve learner accuracy compared to completely supervised methods [33]. In a study by Li et al. [9], a two-view semi-supervised method for review spam detection was created by employing the framework of a co-training algorithm to make use of the large amount of unlabeled reviews available. The co-training algorithm was developed by Blum and Mitchell [34], and is a bootstrapping method that uses a set of labeled data to incrementally apply labels to unlabeled data. It trains 2 classifiers on 2 distinct sets of features and adds the instances most confidently labeled by each classifier to the training set. This effectively allows large datasets to be generated and used for classification, reducing the demand to manually produce labeled training instances. A modified version of the co-training algorithm that only adds instances that were assigned the same label by both classifiers was also proposed. Their dataset was generated with the assistance of students who manually labeled 6000 reviews collected from Epinions.com, 1394 of which were labeled as review spam. Four groups of review centric features were created: content, sentiment, product and metadata. Another two groups of reviewer centric features were created: profile and behavioral.

In order to use the two-view method for adding unlabeled instances to the training set, classifiers were trained on each set of features (i.e., one with review centric features and another with reviewer centric ones). Note that these 2 classifiers are only used to add instances to the labeled data and the final classifier is trained using all available features, both review centric and reviewer centric. Experiments were conducted using Naïve Bayes, Logistic Regression and SVM with 10-fold cross validation, and it was found that Naïve Bayes was the best performer, so all additional work was performed with Naïve Bayes. They observed that using the co-training semi-supervised method, they were able to obtain an F-Score of .609, which was higher than the 0.583 they obtained when not including any unlabeled data. Further, it was observed that by using their co-training with agreement modification, they were able to raise this value to 0.631. While these F-Scores appear low, it is hard to compare them with the performance from other studies as they used their own dataset. The results do seem to indicate that this type of semi-supervised learning may indeed help in the area of review spam detection and demands further study with additional datasets.

PU-Learning is a second type of semi-supervised learning approach, developed by Liu et al. [35], to learn from a few positive examples and a set of unlabeled data. Montes-y-Gómez and Rosso adapt this approach for review spam detection in their work "Using PU-Learning to Detect Deceptive Opinion Spam" [36]. PU-learning is an iterative method which tries to identify a set of reliably negative instances in the unlabeled data. The model is trained and evaluated using all of the unlabeled data as the negative class and any instances that are classified as positive are removed. The process is repeated until some stop criterion is reached. For evaluation purposes, the dataset generated by Ott et al. [3] was used and the performance was evaluated using F-Measure. Classifiers were trained using both Naïve Bayes and SVM as learners. PU-learning achieved an F-measure of 83.7 % with NB, using only 100 positive examples. While this is better than the results achieved using 6000 labeled instances and

co-training by Li F. et al. [9], it is difficult to make a conclusive statement as the methods use different datasets and, as previously discussed, the dataset created by Ott et al. may not provide an accurate indication of real world performance.

Although there is little research in the area of using semi-supervised learning for review spam detection, results obtained using this approach are promising and with additional research, may yield better performance than supervised learning while reducing the need to generate large labeled datasets.

Reviewer centric review spam detection

We mentioned earlier that recognizing reviewers who are writing fake reviews is important in the effort to detect review spam. Using reviewer centric features in combination with review centric features may be preferred over a review centric only approach for spam detection. Additionally, gathering behavioral evidence of spammers is easier than identifying review spam [37].

A thorough study of supervised learning approaches for deceptive review detection was conducted by Mukherjee et al. [23]. They studied how well existing research methods work for detecting real-world fake reviews on a commercial website. The authors tested their models using the Amazon Mechanical Turk (AMT) synthetic fake reviews dataset on a real-world fake reviews dataset procured from Yelp. In this study, they found similar results to previous studies, confirming that using n-gram features performs well on the AMT dataset, however, when used with the real world Yelp dataset it performed significantly worse. They observed that using behavioral features yields higher performance than linguistic features alone on the real world Yelp dataset. Three different features sets were used in the experiment: LIWC, POS and bigrams. In addition, feature selection using Information Gain (IG) was applied to select the top 1 and 2 % features. One of the main conclusions of the study was that the synthetic reviews are not necessarily representative of what is found in real world review spam. Additionally, they observed that using the abnormal behavioral features (i.e., higher percentage of positive reviews, high number of reviews, average review length, etc.) yields better results than the n-gram features in these more realistic datasets. The results of a 5-fold cross validation experiment with an SVM classifier using bigram and POS features resulted in an accuracy of 68.1 % for the real-world fake reviews. This is far lower than the 90 % reported by Ott et al. when evaluating their model on synthetic data. From this, it appears that that using AMT, one cannot effectively generate fake reviews consistent with real-world fake reviews, or at least consistent with the types of reviews that Yelp filters. The addition of behavioral features increases their accuracy to 86.1 % on Yelp's filtered reviews dataset. Feature selection was found to offer no improvement to classification performance, and actually decreased performance slightly; however, only a single combination feature selection technique, learner and performance metric was considered.

In a later study, Mukherjee et al. [14] confirmed that the writers of review spam have different behaviors than truthful reviewers in a set of Amazon reviews as well. Jindal et al. [8] also studied the impact of reviewer centric features on review spam detection. They identified unusual review patterns and reviewer behaviors that were highly correlated with spam review activity. They found unexpected rules and rule

groups using Class Association Rules (CAR), which proposes unexpectedness measures after a set of expectations has been defined. These unexpected rules and rule groups represent the unusual behaviors of spam reviewers, which in turn allow for identification of review spam activity. This technique itself is generic and can be applied to solve a variety of problems due to its domain independence.

A novel technique for detecting review spammers was proposed by Fei. et al. [22], where they exploit the "bursty" nature of reviews generated by spammers to identify review spam. Bursty reviews are reviews that suddenly become popular and receive great attention from reviewers within a certain time period or certain area. The reviews and reviewers in those situations become suspicious as review spam and review spammer respectively. For burst detection, the authors used Kernel Density Estimation (KDE) techniques to detect review bursts. KDE is a technique closely related to histograms, which has attributes that allow it to asymptotically converge to any density function. Behavioral features for spammers were created that combined the spammers' behaviors with the features of review bursts. In addition, these features can be used in conjunction with review spam features in a hybrid approach to improve the classification results. The features listed below are examples of the features used in this study.

Ratio of Amazon Verified Purchase (RAVP)

This feature is the number of the Amazon verified purchases divided by the number of total reviews written by this user. Because verified purchase reviews most likely reflect a genuine review, a reviewer with a higher RAVP is considered more trustworthy.

Rating Deviation (RD)

This feature measures the average deviation of a reviewer's reviews. Since the expected behavior of a reviewer is to give similar ratings as other users gave for the same product, spammers may exhibit a higher divergence in their rating behavior.

Burst Review Ratio (BRR)

This value is computed as the ratio of a reviewer's reviews that occur in bursts to the total number of reviews that he/she wrote.

Review Content Similarity (RCS)

The average pairwise cosine similarity of all of a reviewer's reviews. Higher scores may be an indication of a possible spammer.

Reviewer Burstiness (RB)

This measures the amount of reviews that occur in both the reviewer's and product's bursts. The more that this occurs, the more likely the reviewer is a spammer.

A Markov Random Field (MRF) model engaged with a Loopy Belief Propagation algorithm was used to identify the spam reviewers in their proposed model. The dataset produced by Jindal et al. [21] was used for training and evaluation. Unigrams features were used with SVM to classify the reviews for evaluation purposes, but not used in the main model. Using only reviewer centric features Fei et al. achieved an F-score of 75.4 % for burst reviews, and 68.7 % for all reviews. Earlier results by Jindal et al. [21]

indicate similar performance can be achieved using text based features; however work by Mukherjee et al. [14] shows that classifiers benefit from using both review centric and reviewer centric features.

Comparative analysis and suggestions

When developing a new review spam detection framework, it is important to understand what approaches and techniques have been used in prior studies. In previous sections, we presented an overview of machine learning techniques that have been used in the review spam domain and some of the important results of these studies. As this domain is young, relatively few studies on machine learning techniques and review spam detection have been conducted.

Based on our survey, most of the previous studies have focused on supervised learning techniques. However, in order to use supervised learning, one must have a labeled dataset, which can be difficult (if not impossible) to acquire in the area of review spam. From the literature we discussed, it can be observed that most of the available datasets used in the previous studies are synthetically created, most likely due to the lack of review spam examples and the difficulty of labeling them [19]. Building and evaluating classifiers based on these synthetic datasets can be problematic, as it has been observed that they are not necessarily representative of real world review spam. For example, when using the same framework to evaluate the artificial AMT dataset used in [3, 12, 25] and Yelp's filtered reviews dataset, the extracted features and results differed greatly, especially when using n-gram text features [23]. Comparing classification performance across these datasets shows that when evaluated on the synthetic review dataset, the classifier achieved an accuracy of 87 %, but while using Yelp's reviews only achieved 65 % accuracy. This 22 % drop in accuracy implies that synthetically created reviews have different distinguishing features than real-life fake reviews, and that the reviews produced by AMT do not accurately reflect real world spam reviews.

Feature engineering can have a significant impact on classifier performance. Different studies have used the same datasets, learners, and performance metrics but achieved different results due to different feature engineering methods; [3] and [25] or [23] and [11] are examples. Table 8 reports the performance for some of the studies discussed in this paper and what types of features were used to achieve that value. In studying the various sets of features used in the literature, one of the most notable conclusions is that performance increases through combining multiple types of features, and that using the most relevant and expressive features can make a predictive model more robust [25]. Jindal et al. [21] found that adding additional features (both review centric and reviewer centric) to text features improved performance. It can also be observed in Table 8 that augmenting bigrams with LIWC yields a small performance improvement [3]. Several experiments used the same datasets (built by Ott et al. [3] using AMT) and show that for this dataset, the highest performance is achieved using bigrams and LIWC [3, 11, 12]. As other studies are using unique datasets, or datasets that have been in some way altered, it is difficult to directly compare their results.

Although there are a large number of machine learning algorithms (learners) available, current research using supervised learning methods has been, for the most part,

limited to three learners: Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Machine (SVM). While SVM generally offered the best performance; it is occasionally beaten by NB or LR, and not compared to many other available learners, thus it cannot be considered the best learner. The best learner found by each study is shown in Table 9, but should not be considered conclusive due to the experiments not thoroughly studying multiple learners. Future research should test multiple learners across multiple datasets using many different feature engineering methods.

To the best of our knowledge, methods and tools for learning from Big Data have not been used in the literature even though real world datasets of only a single site (such as Trip Advisor) can contain upwards of 200 million reviews[5]. New reviews are constantly being added to large repositories of reviews across various websites at a high rate, over 1.5 million per month in the case of Yelp[4]. Consequently, distributed and streaming applications of machine learning algorithms

Table 9 Comparison of previous works and results for review spam detection along with the relative complexity of the approach (including feature extraction and learning methodology)

Paper	Dataset	Features used	Learner	Performance metric	Score	Method complexity
[20]	5.8 million reviews written by 2.14 reviewers crawled from amazon website	Review and reviewer features	LR	AUC	78 %	Low
[21]	5.8 million reviews written by 2.14 reviewers crawled from amazon website	Features of the review, reviewer and product characteristics	LR	AUC	78 %	Medium
[21]	5.8 million reviews written by 2.14 reviewers crawled from amazon website	Text features	LR	AUC	63 %	Low
[9]	6000 reviews from Epinions	Review and reviewer features	NB with Co-training	F-Score	0.631	High
[3]	Hotels through Amazon Mechanical Turk (AMT) by Ott et al.	Bigrams	SVM	Accuracy	89.6 %	Low
[3]	Hotels through Amazon Mechanical Turk (AMT) by Ott et al.	LIWC + Bigrams	SVM	Accuracy	89.8 %	Medium
[25]	Hotels through Amazon Mechanical Turk (AMT) by Ott et al. + gathered 400 deceptive hotel and doctor reviews from domain experts	LIWC + POS + Unigram	SAGE	Accuracy	65 %	High
[23]	Yelp's real-life data	Behavioral features combined with the bigram features	SVM	Accuracy	86.1 %	Medium
[11]	Hotels through Amazon Mechanical Turk (AMT) by Ott et al.	Stylometric features	SVM	F-measure	84 %	Low
[12]	Hotels through Amazon Mechanical Turk (AMT) by Ott et al.	n-gram features	SVM	Accuracy	86 %	Low
[1]	Dataset collected from amazon.com	Syntactical, lexical, and stylistic features	SLM	AUC	.9986	High
[24]	Their own crawled Arabic reviews from tripadvisor.com, booking.com, and agoda.ae	Review and reviewer features	NB	F-measure	.9959	Low

across these datasets are of interest as traditional machine learning tools, such as R or Weka, cannot scale to datasets of this size. Tools such as Mahout[11], Spark (MLlib)[12], H_2O[13], and SAMOA[14] should be explored to effectively model the large corpus of online reviews which exist in the real world [38]. Mahout has been used for large-scale recommendation systems [39], which would be useful to apply to review spam detection, as reviewers may be related to each other on different review websites. MLlib and SAMOA can perform large-scale online learning, where machine learning models are trained and tuned as new data flows in. This is especially desirable in the field of review spam detection, as reviews are constantly being added to the corpus. SAMOA has been used to analyze live Twitter streams [40], which involves similar text processing that can be applied to online reviews.

Current research has largely ignored feature selection techniques in their experiments, even when using text features, which can potentially lead to highly dimensional feature sets. The experiment by Mukherjee et al. [23] is a notable exception, as they used Information Gain (IG) to perform feature selection of top 1 and 2 % of features. Though they found this had no impact on classifier performance, we believe that using feature selection techniques can potentially improve performance based on results from other domains. Feature selection also has the benefit of reducing the computational costs associated with training a classifier. This is highly desirable as review spam detection is a big data domain and datasets may have a very large number of instances and features. In order to ascertain the impact of feature selection, additional techniques should be tested while considering different features, feature subset sizes and datasets.

In addition, current research has ignored the use of ensemble learning techniques, such Bagging or Boosting, to obtain better predictive performance than using the traditional learning algorithms. These techniques are especially useful for improving performance on noisy or imbalanced data [41, 42]. Noisy data is data with inaccuracies or, "noise", in either the features or class attributes. For example, training data may contain review spam instances that have been mislabeled as true reviews or vice versa [43]. As classification performance on synthetic review datasets has shown to be a poor indicator of performance on real world data, it is beneficial to use real world data. Unfortunately it is difficult to accurately label training data. As seen in the study by Ott et al. [3], human judges have difficulty in accurately discriminating between, and thus labeling, spam and non-spam reviews. It is likely any labeled training data from real world sources would contain mislabeled instances. Due to this, ensemble techniques could be highly beneficial in this domain to mitigate the negative impact of noisy data.

Finally, there are a massive number of online reviews, and fake reviews are usually less frequent than truthful ones, resulting in highly imbalanced datasets [44]. Class imbalance can adversely affect classifier performance as the majority class may be favored, and must be taken into consideration when training a model. Two works have considered the class imbalance problem in this domain, [24] and [44]. Both used random undersampling and random oversampling to overcome imbalanced distributions and have promising but inconclusive results. Ensemble techniques can be used alongside, or in place of, data sampling as they have been shown to be more robust to the effects of class imbalance than single classifiers

[41], but have yet to be used to address imbalanced data in this domain. Future work should include further investigation of the role class imbalance in review spam data as well as mitigating its effects using ensemble learners and sampling techniques.

Conclusion

In recent years, review spam detection has received significant attention in both business and academia due to the potential impact fake reviews can have on consumer behavior and purchasing decisions. This survey covers machine learning techniques and approaches that have been proposed for the detection of online spam reviews. Supervised learning is the most frequent machine learning approach for performing review spam detection; however, obtaining labeled reviews for training is difficult and manual identification of fake reviews has poor accuracy. This has led to many experiments using synthetic or small datasets. Features extracted from review text (e.g., bag of words, POS tags) are often used to train spam detection classifiers. An alternative approach is to extract features related to the metadata of the review, or features associated with the behavior of users who write the reviews. Disparities in performance of classifiers on different datasets may indicate that review spam detection may benefit from additional cross domain experiments to help develop more robust classifiers. Multiple experiments have shown that incorporating multiple types of features can result in higher classifier performance than using any single type of feature.

One of the most notable observations of current research is that experiments should use real world data if possible. Despite being used in many studies, synthetic or artificially generated datasets have been shown to give a poor indication of performance on real world data [23]. As it is difficult to procure accurately labeled real-world datasets, unsupervised and semi-supervised methods are of interest. While unsupervised and semi-supervised methods are currently unable to match the performance of supervised learning methods, research is limited and results are inconclusive, warranting further investigation. A possibility for a less labor-intensive means of generating labeled training data is to find and label duplicate reviews as spam. Multiple studies have shown duplication, or near duplication, of review content is a strong indicator of review spam. Another data related concern is that real world data may be highly class imbalanced, as there are currently many more truthful than fake reviews online. This could be addressed through data sampling and ensemble learning techniques. A final concern related to quality of data is the presence of noise, particularly class noise due to mislabeled instances. Ensemble methods, and experiments with different levels of class noise, could be used to evaluate the impact of noise on performance and how its effects may be reduced.

The studies discussed in this paper have primarily focused in the area of feature engineering, but which combination of features is best remains unclear. Research by Jindal et al. [20, 21] shows that the addition of reviewer centric features yields higher classifier performance than the use of n-gram features alone, and other experiments support this conclusion [3, 9, 23]. The best observed performance was achieved by combining text and non-text features. Reviewer centric features have also been demonstrated to be important for accurate detection of review spam as seen in [9, 20, 21, 23, 24]. Despite many studies focusing on feature engineering, it is not possible to identify

the best types of features since the experiments make use of different datasets; however, it has been shown that there is no silver bullet for review spam detection and multiple types of features are needed. Future work should evaluate different feature engineering methods across multiple datasets to determine which types of features are most useful for online review spam detection.

As review text is an important source of information and tens of thousands of text features can easily be generated based on this text, high dimensionality can be an issue. Additionally, millions of reviews are available to be used to train classifiers, and training classifiers from a large, highly dimensional dataset is computationally expensive and potentially impractical. Despite this, feature selection techniques have received little attention. Many experiments have avoided this issue by extracting only a small number of features, avoiding the use of n-grams, or by limiting number of features through alternative means such as using term frequencies to determine what n-grams are included as features. Further work needs to be conducted to establish how many features are required and what types of features are the most beneficial. Feature selection should not be considered optional when training a classifier in a big data domain with potential for high feature dimensionality. Additionally, we could find no studies that incorporated distributed or streaming implementations for learning from Big Data into their spam detection frameworks.

Endnotes

[1]http://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/03782.html

[2]http://consumerist.com/2010/04/14/how-you-spot-fake-online-reviews/

[3]http://www.moneytalksnews.com/2011/07/25/3-tips-for-spotting-fake-product-reviews-%E2%80%93-from-someone-who-wrote-them/

[4]http://phx.corporate-ir.net/External.File?item=UGFyZW50SUQ9MjY5MzkwfENoaWxkSUQ9LTF8VHlwZT0z&t=1

[5]http://www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html

[6]http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[7]http://www.liwc.net/

[8]http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

[9]http://www.std.org/~msm/common/clustering.html

[10]https://www.mturk.com/mturk/

[11]http://mahout.apache.org/

[12]https://spark.apache.org/mllib/

[13]http://h2o.ai/

[14]http://samoa-project.net/

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
HA began the literature review and manuscript. MC performed the primary literature review and analysis, and updated the manuscript. JDP and ANR performed additional analysis, modified the manuscript, and helped organize the focus. TMK introduced this topic to HA, MC, JDP, and ANR, and coordinated the authors to complete and finalize this work. All authors read and approved the final manuscript.

Received: 8 June 2015 Accepted: 24 August 2015
Published online: 05 October 2015

References

1. Lau RY, Liao SY, Kwok RCW, Xu K, Xia Y, Li Y (2011) Text mining and probabilistic language modeling for online review spam detecting. ACM Trans Manage Inf Syst 2(4):1–30
2. Dixit S, Agrawal AJ (2013) Survey on review spam detection. Int J Comput Commun Technol ISSN (PRINT) 4:0975–7449
3. Ott M, Choi Y, Cardie C, Hancock JT (2011) Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 309–319). Association for Computational Linguistics
4. López V, del Río S, Benítez JM, Herrera F (2015) Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. Fuzzy Sets Syst 258:5–38
5. Bing L (2008) Web Data Mining. Book. Springer, Berlin Heidelberg New York
6. Bandakkanavar RV, Ramesh M, Geeta H (2014) A survey on detection of reviews using sentiment classification of methods. IJRITCC 2(2):310–314
7. Abbasi A, Zhang Z, Zimbra D, Chen H, Nunamaker JF Jr (2010) Detecting fake websites: the contribution of statistical learning theory. MIS Q 34(3):435–461
8. Jindal N, Liu B, Lim EP (2010) Finding unusual review patterns using unexpected rules. In: Proceedings of the 19th ACM international conference on Information and knowledge management. (pp. 1549–1552). ACM, Toronto, ON, Canada
9. Li F, Huang M, Yang Y, Zhu X (2011) Learning to identify review spam. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol 22, No. 3., p 2488
10. Mukherjee A, Liu B, Glance N (2012) Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st international conference on World Wide Web. (pp. 191–200). ACM, Lyon, France
11. Shojaee S, Murad MAA, Bin Azman A, Sharef NM, Nadali S (2013) Detecting deceptive reviews using lexical and syntactic features. In: Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on (pp. 53–58). IEEE, Serdang, Malaysia
12. Ott M, Cardie C, Hancock JT (2013) Negative Deceptive Opinion Spam. In: HLT-NAACL., pp 497–501
13. Qian T, Liu B (2013) Identifying Multiple Userids of the Same Author. In: EMNLP., pp 1124–1135
14. Mukherjee A, Kumar A, Liu B, Wang J, Hsu M, Castellanos M, Ghosh R (2013) Spotting opinion spammers using behavioral footprints. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 632–640). Chicago, ACM.
15. Feng S, Xing L, Gogar A, Choi Y (2012) Distributional footprints of deceptive product reviews. ICWSM 12:98–105
16. Xie S, Wang G, Lin S, Yu PS (2012) Review spam detection via temporal pattern discovery. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 823–831). ACM, Beijing, China
17. Wang G, Xie S, Liu B, Yu PS (2012) Identify online store review spammers via social review graph. ACM Transactions on Intelligent Systems and Technology (TIST) 3(4):61
18. Wang G, Xie S, Liu B, Yu PS (2011) Review graph based online store review spammer detection. In: Data mining (icdm), 2011 ieee 11th international conference on (pp. 1242–1247). IEEE, Vancouver, Canada
19. Morales A, Sun H, Yan X (2013) Synthetic review spamming and defense. In: Proceedings of the 22nd international conference on World Wide Web companion (pp. 155–156). International World Wide Web Conferences Steering Committee, Rio de Janeiro, Brazil
20. Jindal N, Liu B (2007) Review spam detection. In: Proceedings of the 16th international conference on World Wide Web (pp. 1189–1190). ACM, Lyon, France
21. Jindal N, Liu B (2008) Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 219–230). ACM, Stanford, CA
22. Fei G, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R (2013) Exploiting Burstiness in reviews for review spammer detection. ICWSM 13:175–184
23. Mukherjee A, Venkataraman V, Liu B, Glance NS (2013) What yelp fake review filter might be doing? Boston, In ICWSM.
24. Hammad ASA (2013) An Approach for Detecting Spam in Arabic Opinion Reviews. Doctoral dissertation, Islamic University of Gaza
25. Li J, Ott M, Cardie C, Hovy E (2014) Towards a general rule for identifying deceptive opinion spam. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1566–1576, Baltimore, Maryland, USA, June 23-25 2014. ACL
26. Part of Speech Tagging (POS). http://en.wikipedia.org/wiki/Part-of-speech_tagging
27. Mayzlin D, Dover Y, Chevalier JA (2012) Promotional reviews: An empirical investigation of online review manipulation (No. w18340). National Bureau of Economic Research, Nashville, TN
28. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
29. Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ (2007) The development and psychometric properties of LIWC2007
30. Eisenstein J, Ahmed A, Xing EP (2011) Sparse additive generative models of text. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11)., pp 1041–1048
31. Abbasi A, Chen H, Nunamaker JF (2008) Stylometric identification in electronic markets: Scalability and robustness. J Manage Inf Syst 25(1):49–78
32. Mukherjee A, Venkataraman V, Liu B, Glance N (2013) Fake review detection: Classification and analysis of real and pseudo reviews. Technical Report UIC-CS-2013-03, University of Illinois, Chicago
33. Chapelle O, Schölkopf B, Zien A (2006) Semi-supervised learning. Vol. 2. Cambridge: MIT press.
34. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory (pp. 92–100). ACM, Madison, WI
35. Liu B, Dai Y, Li X, Lee WS, Yu PS (2003) Building text classifiers using positive and unlabeled examples. In: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 179–186). Melbourne, Florida, IEEE
36. Hernández D, Guzmán R, Móntes y Gomez M, Rosso P (2013) Using PU-learning to detect deceptive opinion spam. In: Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis., pp 38–45

37. Lim EP, Nguyen VA, Jindal N, Liu B, Lauw HW (2010) Detecting product review spammers using rating behaviors. In: Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 939–948). ACM, Toronto, CA
38. Richter AN, Khoshgoftaar TM, Landset S, Hasanin T. A Multi-Dimensional Comparison of Toolkits for Machine Learning with Big Data. IEEE (in press)
39. Sumbaly R, Kreps J, Shah S (2013) The big data ecosystem at LinkedIn. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data on (pp.1125- 1134). SIGMOD, NY, NY
40. Kourtellis N, Morales GDF, Bonchi F, De G, Morales F (2014) Scalable online betweenness centrality in evolving graphs. CoRR, abs/1401.6981
41. Dietterich TG (2000) Ensemble methods in machine learning. In: Multiple classifier systems. Springer, Berlin Heidelberg, pp 1–15
42. Khoshgoftaar TM, Van Hulse J, Napolitano A (2011) Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans Syst Man Cybern A Syst Hum 41(3):552–568
43. Van Hulse J, Khoshgoftaar T (2009) Knowledge discovery from imbalanced and noisy data. Data Knowl Eng 68(12):1513–1542
44. Al Najada H, Zhu X (2014) iSRD: Spam review detection with imbalanced data distributions. In: Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on (pp. 553–560). IEEE, San Francisco, CA