Towards a General Rule for Identifying Deceptive Opinion Spam

Jiwei Li[1], **Myle Ott**[2], **Claire Cardie**[2], **Eduard Hovy**[1]
[1]Language Technology Institute, Carnegie Mellon University, Pittsburgh, P.A. 15213, USA
[2]Department of Computer Science, Cornell University, Ithaca, N.Y., 14853, USA
bdlijiwei@gmail.com, myleott@cs.cornell.edu
cardie@cs.cornell.edu, ehovy@andrew.cmu.edu

Abstract

Consumers' purchase decisions are increasingly influenced by user-generated online reviews. Accordingly, there has been growing concern about the potential for posting deceptive opinion spam—fictitious reviews that have been deliberately written to sound authentic, to deceive the reader. In this paper, we explore generalized approaches for identifying online deceptive opinion spam based on a new gold standard dataset, which is comprised of data from three different domains (i.e. Hotel, Restaurant, Doctor), each of which contains three types of reviews, i.e. *customer generated truthful reviews*, *Turker generated deceptive reviews* and *employee (domain-expert) generated deceptive reviews*. Our approach tries to capture the general difference of language usage between deceptive and truthful reviews, which we hope will help customers when making purchase decisions and review portal operators, such as TripAdvisor or Yelp, investigate possible fraudulent activity on their sites.[1]

1 Introduction

Consumers increasingly rely on user-generated online reviews when making purchase decision (Cone, 2011; Ipsos, 2012). Unfortunately, the ease of posting content to the Web, potentially anonymously, creates opportunities and incentives for unscrupulous businesses to post *deceptive opinion spam*—fictitious reviews that are deliberately written to sound authentic, in order to deceive the reader.[2] Accordingly, there appears to be widespread and growing concern among both businesses and the public about this potential abuse (Meyer, 2009; Miller, 2009; Streitfeld, 2012; Topping, 2010; Ott, 2013).

Existing approaches for spam detection are usually focused on developing supervised learning-based algorithms to help users identify deceptive opinion spam, which are highly dependent upon high-quality gold-standard labeled data (Jindal and Liu, 2008; Jindal et al., 2010; Lim et al., 2010; Wang et al., 2011; Wu et al., 2010). Studies in the literature rely on a couple of approaches for obtaining labeled data, which usually fall into two categories. The first relies on the judgements of human annotators (Jindal et al., 2010; Mukherjee et al., 2012). However, recent studies show that deceptive opinion spam is not easily identified by human readers (Ott et al., 2011). An alternative approach, as introduced by Ott et al. (2011), crowdsourced deceptive reviews using Amazon Mechanical Turk.[3] A couple of follow-up works have been introduced based on Ott et al.'s dataset, including estimating prevalence of deception in online reviews (Ott et al., 2012), identification of negative deceptive opinion spam (Ott et al., 2013), and identifying manipulated offerings (Li et al., 2013b).

Despite the advantages of soliciting deceptive gold-standard material from Turkers (it is easy, large-scale, and affordable), it is unclear whether Turkers are representative of the general population that generate fake reviews, or in other words, Ott et al.'s data set may correspond to only one type of online deceptive opinion spam — fake reviews generated by people who have never been to offerings or experienced the entities. Specifically, according to their findings (Ott et al., 2011;

[1] Dataset available by request from the first author.

[2] Manipulating online reviews may also have legal consequences. For example, the Federal Trade Commission (FTC) has updated their guidelines on the use of endorsements and testimonials in advertising to suggest that posting deceptive reviews may be unlawful in the United States (FTC, 2009).

[3] http://www.mturk.com

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1566–1576,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

Li et al., 2013a), truthful hotel reviews encode more spatial details, characterized by terms such as "bathroom" and "location", while deceptive reviews talk about general concepts such as why or with whom they went to the hotel. However, a hotel can instead solicit fake reviews from their employees or customers who possess substantial domain knowledge to write fake reviews and encode more spatial details in their lies. Indeed, cases have been reported where hotel owners bribe guests in return for good reviews on TripAdvisor[4], or companies ordered employees to pretend they were satisfied customers and write glowing reviews of its face-lift procedure on Web sites.[5] The domain knowledge possessed by domain experts enables them to craft reviews that are much more difficult for classifiers to detect, compared to the crowdsourced fake reviews.

Additionally, existing supervised algorithms in the literature are usually narrowed to one specific domain and heavily rely on domain-specific vocabulary. For example, classifiers assign high weights to domain-specific terms such as "hotel", "rooms", or even the name of the hotels such as "Hilton" when trained on reviews on hotels. It is unclear whether these classifiers will perform well at detecting deception in other domains, e.g., Restaurant or Doctor reviews. Even in a single domain, e.g., Hotel, classifiers trained from reviews of one city (e.g., Chicago) may not be effective if directly applied to reviews from other cities (e.g., New York City) (Li et al., 2013b). In the examples in Table 1, we trained a linear SVM classifier on Ott's Chicago-hotel dataset on *unigram* features and tested it on a couple of different domains (the details of data acquisition are illustrated in Section 3). Good performance is obtained on Chicago-hotel reviews (Ott et al., 2011), but not as good on New York City ones. The performance is reasonable in Restaurant reviews due to the many shared properties among restaurants and hotels, but suffers in Doctor settings.

In this paper, we try to obtain a deeper understanding of the general nature of deceptive opinion spam. One contribution of the work presented here is the creation of the cross-domain (i.e., Hotel, Restaurant and Doctor) gold-standard dataset.

	Accuracy	Precision	Recall	F1
NYC-Hotel	0.799	0.794	0.758	0.766
Chicago-Restaurant	0.785	0.813	0.742	0.778
Doctor	0.550	0.537	0.725	0.617

Table 1: SVM performance on datasets for a classifier trained on Chicago hotel review based on Unigram feature.

In contrast to existing work (Ott et al., 2011; Li et al., 2013b), our new gold standard includes three types of reviews: *domain expert deceptive opinion spam* (*Employee*), *crowdsourced deceptive opinion spam* (*Turker*), and *truthful Customer reviews* (*Customer*). In addition, some of domains contain both positive (**P**) and negative (**N**) reviews.[6]

To explore the general rule of deceptive opinion spam, we extended SAGE Model (Eisenstein et al., 2011), 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) in the text collection. We find that more general features, such as LIWC and POS, are more robust when modeled using SAGE, compared with just bag-of-words.

We additionally make theoretical contributions that may shed light on a longstanding debate in the literature about deception. For example, in contrast to existing findings that highlight the lack of spatial detail in deceptive reviews (Ott et al., 2011; Li et al., 2013b), we find that a lack of spatial detail may not be a universal cue to deception, since it does not apply to fake reviews written by domain experts. Instead, our finding suggest that other linguistic features may offer more robust cues to deceptive opinion spam, such as overly highlighted sentiment in the review or the overuse of first-person singular pronouns.

The rest of this paper is organized as follows. In Section 2, we briefly go over related work. We describe the creation of our data set in Section 3 and present our model in Section 4. Experimental results are shown in Section 5. We present analysis of general cues to deception in Section 6 and conclude this paper in Section 7.

[4] http://www.dailymail.co.uk/travel/article-2013391/Tripadvisor-Hotel-owners-bribe-guests-return-good-reviews.html

[5] `http://www.nytimes.com/2009/07/15/technology/internet/15lift.html?_r=0`

[6] For example, a hotel manager could hire people to write positive reviews to increase the reputation of his own hotel or post negative ones to degrade his competitors. Identifying positive/negative opinion spam is explored in (Ott et al., 2011; Ott et al., 2013)

2 Related Work

Spam has been historically studied in the contexts of Web text (Gyöngyi et al., 2004; Ntoulas et al., 2006) or email (Drucker et al., 1999). Recently there has been increasing concern about deceptive opinion spam (Jindal and Liu, 2008; Ott et al., 2011; Wu et al., 2010; Mukherjee et al., 2013b; Wang et al., 2012).

Jindal and Liu (2008) first studied the deceptive opinion problem and trained models using features based on the review text, reviewer, and product to identify duplicate opinions, i.e., opinions that appear more than once in the corpus with similar contexts. Wu et al. (2010) propose an alternative strategy to detect deceptive opinion spam in the absence of a gold standard. Yoo and Gretzel (2009) gathered 40 truthful and 42 deceptive hotel reviews and manually compare the linguistic differences between them. Ott et al. created a gold-standard collection by employing Turkers to write fake reviews, and follow-up research was based on their data (Ott et al., 2012; Ott et al., 2013; Li et al., 2013b; Feng and Hirst, 2013). For example, Song et al. (2012) looked into syntactic features from Context Free Grammar parse trees to improve the classifier performance. A step further, Feng and Hirst (2013) make use of degree of *compatibility* between the personal experiment and a collection of reference reviews about the same product rather than simple textual features.

In addition to exploring text or linguistic features in deception, some existing work looks into customers' behavior to identify deception (Mukherjee et al., 2013a). For example, Mukherjee et al. (2011; 2012) delved into group behavior to identify group of reviewers who work collaboratively to write fake reviews. Qian and Liu (2013) identified multiple user IDs that are generated by the same author, as these authors are more likely to generate deceptive reviews.

In the psychological literature, researchers have looked into possible linguistic cues to deception (Newman et al., 2003), such as *decreased spatial detail*, which is consistent with theories of reality monitoring (Johnson and Raye, 1981), increased negative emotion terms (Newman et al., 2003), or the writing style difference between informative (truthful) and imaginative (deceptive) writings in (Rayson et al., 2001). The former typically consists of more nouns, adjectives, prepositions, determiners, and coordinating conjunctions, while the latter consists of more verbs, adverbs, pronouns, and pre-determiners.

SAGE (Sparse Additive Generative Model): SAGE is an generative bayesian approach introduced by Eisenstein et al. (2011), which can be viewed as an combination of topic models (Blei et al., 2003) and generalized additive models (Hastie and Tibshirani, 1990). Unlike other derivatives of topic models, SAGE drops the Dirichlet-multinomial assumption and adopts a Laplacian prior, triggering sparsity in topic-word distribution. The reason why SAGE is tailored for our task is that SAGE constructs multi-faceted latent variable models by simply adding together the component vectors rather than incorporating multiple switching latent variables in multiple facets.

3 Dataset Construction

In this section, we report our efforts to gather gold-standard opinion spam datasets. Our datasets contain the following domains, namely Hotel, Restaurant, and Doctor.

3.1 *Turker* set, using Mechanical Turk

Crowdsourcing services such as AMT greatly facilitate large-scale data annotation and collection efforts. Anyone with basic programming skills can create Human Intelligence Tasks (HITs) and access a marketplace of anonymous online workers (Turkers) willing to complete the tasks. We borrowed some rules used by Ott et al. to create their dataset, such as restricting task to Turkers located in the United States, and who maintain an approval rating of at least 90%.

Hotel-Turker : We directly borrowed datasets from Ott[7] and Li.[8]

Restaurant-Turker : We gathered 20 positive (P) deceptive reviews for each of 10 of the most popular restaurants in Chicago, for a total of 200 positive deceptive restaurant reviews.

Doctor-Turker : We gathered a total number of 200 positive reviews from Turkers.

3.2 *Employee* set, by domain experts

We seek deceptive opinion spam written by people with expert-level domain knowledge. It is not appropriate to use crowdsourcing to obtain this data,

[7] http://myleott.com/op_spam/

[8] http://www.cs.cmu.edu/~jiweil/html/four_city.html

	Turker	*Expert*	*Customer*
Hotel (P/N)	400/400	140/140	400/400
Restaurant (P/N)	200/0	120/0	200/200
Doctor (P/N)	200/0	32/0	200/0

Table 2: Statistics for our dataset.

so instead we solicit reviews written by employees in each domain.

Hotel-Employee: We asked two hotel employees from each of seven hotels (14 employees total) each to write 10 deceptive positive-sentiment reviews of their own hotel, and 10 deceptive negative-sentiment reviews of their biggest local competitor's hotel. In total, we obtained 280 deceptive reviews of 14 hotels, including a balanced mix of positive- and negative-sentiment reviews.

Restaurant-Employee: We asked employees from selected restaurants (a waiter/waitress or cook) to each write positive-sentiment reviews of their restaurant.

Doctor-Employee: We asked real doctors to write positive fake reviews about themselves. In total we obtained 32 reviews from 15 doctors.

3.3 *Customer* set from Actual Customers

Hotel-Customer: We borrowed from Ott et al.'s dataset.

Restaurant/Doctor-Customer: We solicited data by matching a set of truthful reviews as Ott et al. did in collecting truthful hotel reviews.

3.4 Summary for Data Creation

Statistics for our data set is presented in Table 2. Due to the difficulty in obtaining gold-standard data in the literature, there is no doubt that our data set is not perfect. Some parts are missing, some are unbalanced, participants in the survey may not be representative of the general population. However, as far as we know, this is the most comprehensive dataset for deceptive opinion spam so far, and may to some extent shed insights on the nature of online deception.

4 Feature-based Additive Model

In this section, we briefly describe our model. Since mathematics are not the main theme of this paper, we omit the exact details for inference, which can be found in (Eisenstein et al., 2011).

Before describing the model in detail, we note the following advantages of the SAGE model, and our reasons for using it in this paper:

1. the "additive" nature of SAGE allows a better understanding of which features contribute most to each type of deceptive review and how much each such feature contributes to the final decision jointly. If we instead use SVM, for example, we would have to train classifiers one by one (due to the distinct features from different sources) to draw conclusions regarding the differences between Turker vs Expert vs truthful reviews, positive expert vs negative expert reviews, or reviews from different domains. This would not only become intractable, but would make the conclusions less clear.

2. For cross-domain classification task, standard machine learning approaches may suffer due to domain-specific properties (See Section 5.2).

4.1 Model

In SAGE, each term w is drawn from a distribution proportional to $\exp(m^{(w)} + \eta_{y_d}^{(T)(w)} + \eta_{z_n}^{(A)(w)} + \eta_{y_d,z_n}^{(I)(w)})$, where $m^{(w)}$ is the observed background term frequency, η_{y_d}, η_{z_n} and η_{y_d,z_n} denote the log frequency deviation representing topic z_n, facet y_d, and the second-order interaction part respectively. Superscripts T, A and I respectively denote the index of the topic, facet, and second-order interaction. In our task, we adapt the SAGE model as follows:

$$\begin{aligned} Y = \{&y_{Sentiment} \in \{\text{positive, negative}\},\\ &y_{Domain} \in \{\text{hotel, restaurant, doctor}\},\\ &y_{Source} \in \{\text{employee, turker, customer}\}\} \end{aligned}$$

We model three η's, one for each type of y. Let i, j, k denote the index of the different types of y, so that each term w is drawn as follows:

$$\begin{aligned} P(w|i,j,k) \propto \exp(m^{(w)} + \eta_{y_{Sentiment}}^{(i)(w)} \\ + \eta_{y_{Domain}}^{(j)(w)} + \eta_{y_{Scource}}^{(k)(w)} + \textit{higher order}) \end{aligned}$$

where the *higher order* parts denote the interactions between different facets.

In our approach each document-level feature f is drawn from the following distribution:

$$\begin{aligned} P(f|i,j,k) \propto \exp(m^{(f)} + \eta_{y_{Sentiment}}^{(i)(f)} + \eta_{y_{Domain}}^{(j)(f)} \\ + \eta_{y_{Scource}}^{(k)(f)} + \textit{higher order}) \end{aligned} \tag{1}$$

where $m^{(f)}$ can be interpreted as the background value of feature f. For each review d, the probability that it is drawn from facets with index i, j, k is as follows:

$$P(d|i,j,k) = \prod_{f \in d} P(f|i,j,k) \prod_{w \in d} P(w|i,j,k) \quad (2)$$

In the training process, parameters $\eta_y^{(w)}$ and $\eta_y^{(f)}$ are to be learned by maximizing the posterior distribution following the original SAGE training procedure. For prediction, we estimate y_{Source} for each document given all or part of $\eta_y^{(w)}$ and $\eta_y^{(f)}$ as follows:

$$y_{Source} = \underset{y'_{Source}}{\operatorname{argmax}} P(d|y'_{Source}, y_{Sentiment}, y_{Domain}),$$

where we assume $y_{Sentiment}$ and y_{Domain} are given for each document d. Note that we assume conditional independence between features and words given y, similar to other topic models (Blei et al., 2003). Notably, our revised SAGE model degenerates into a model similar to Generalized Additive Model (Hastie and Tibshirani, 1990) when word features are not considered.

5 Experiments

In this section, we report our experimental results. We first restrict experiments to the within-domain task and see what features most characterize the deceptive reviews, and how. We later extend it to cross domains to explore a more general classifier of deceptive opinion spam.

5.1 Intra-Domain Classification

We explore the effect of both domain experts and crowdsourcing workers on intra-domain deception. Specifically, we reframe it as a **intra-domain multi-class classification task**, where given the labeled training data from one domain, we learn a classifier to classify reviews according to their source, i.e., *Employee*, *Turker* and *Customer*. Since the machine learning classifier is trained and tested within the same domain, $\eta_{y_{Domain}}^{(j)(w)}$ and $\eta_{y_{Domain}}^{(i)(f)}$ are not considered here.

We use a *One-Versus-Rest* (OvR) scheme, in which we train m classifiers using SAGE, such that each classifier f_i, for $i \in [1, m]$, is trained to distinguish between class i on the one hand, and all classes except i on the other. To make an m-way decision, we then choose the class c with the most confident prediction. OvR approaches have been shown to produce state-of-art performance compared to other multi-class approaches such as *Multinomial Naive Bayes* or *One-Versus-One* classification scheme. We train the OvR classifier on three sets of features, *LIWC*, *Unigram*, and *POS*.[9]

Multi-class classification results are given at Table 3. We report both OvR performance and the performance of three One-versus-One binary classifiers, trained to distinguish between each pair of classes. In particular, the three-class classifier is around 65% accurate at distinguishing between *Employee*, *Customer*, and *Turker* for each of the domains using Unigram, significantly higher than random guess. We also observe that each of the three *One-versus-One* binary classifications performs significantly better than chance, suggesting that *Employee*, *Customer*, and *Turker* are in fact three different classes. In particular, the two-class classifier is around 0.76 accurate in distinguishing between *Turker* and *Employee* reviews, despite both kinds of reviews being deceptive opinion spam.

Best performance is achieved on Unigram features, constantly outperforming LIWC and POS features in both three-class and two-class settings in the hotel domain. Similar results are observed for restaurant and doctor domains and details are excluded for brevity. This suggests that a universal set of keyword-based deception cues (e.g., LIWC) is not the best approach for Intra-Domain Classification. Similar results were also reported in previous work (Ott et al., 2012; Ott, 2013).

5.2 Cross-domain Classification

In this subsection, we frame our problem as a *domain adaptation task* (Pan and Yang, 2010). Again, we explore 3 feature sets: *LIWC*, *Unigram* and *POS*. We train a classifier on hotel reviews, and evaluate the performance on other domains. For simplicity, we focus on **truthful** (*Customer*) versus **deceptive** (*Turker*) binary classification rather than a multi-class classification.

We report results from SAGE and SVM[10] in Table 4. We first observe that classifiers trained on hotel reviews apply well in the restaurant domain, which is reasonable due to the many shared prop-

[9] Part-of-speech tags were assigned based on Stanford Parser `http://nlp.stanford.edu/software/lex-parser.shtml`

[10] We use SVMlight (Joachims, 1999) to train our linear SVM classifiers

Domain	Setting	Features		*Customer*		*Employee*		*Turker*	
			A	P	R	P	R	P	R
Hotel	Three-Class	Unigram	0.664	0.678	0.669	0.589	0.610	0.641	0.582
		LIWC	0.602	0.617	0.613	0.541	0.598	0.590	0.511
		POS	0.517	0.532	0.669	0.481	0.479	0.482	0.416
	Customer vs *Turker*	Unigram	0.818	0.812	0.840	-	-	0.820	0.809
		LIWC	0.764	0.774	0.771	-	-	0.723	0.749
		POS	0.729	0.748	0.692	-	-	0.707	0.759
	Customer vs *Employee*	Unigram	0.799	0.832	0.784	0.804	0.820	-	-
		LIWC	0.732	0.746	0.751	0.714	0.722	-	-
		POS	0.728	0.713	0.742	0.707	0.754	-	-
	Employee vs *Turker*	Unigram	0.762	-	-	0.786	0.806	0.826	0.794
		LIWC	0.720	-	-	0.728	0.726	0.698	0.739
		POS	0.701	-	-	0.688	0.710	0.701	0.697
Restaurant	Three-Class	Unigram	0.647	0.692	0.725	0.625	0.648	0.686	0.702
	Customer vs *Turker*		0.817	0.842	0.816	-	-	0.804	0.812
	Customer vs *Employee*		0.785	0.790	0.814	0.769	0.826	-	-
	Employee vs *Turker*		0.774	-	-	0.784	0.804	0.802	0.763
Doctor	*Customer* vs *Turker*		0.745	0.772	0.701	-	-	0.752	0.718

Table 3: Within-domain multi-class classifier performance.

Model	Features	Domain	A	P	R	F1	Domain	A	P	R	F1
SVM	Unigram	Restaurant	0.785	0.813	0.742	0.778	Doctor	0.550	0.537	0.725	0.617
	LIWC	Restaurant	0.745	0.692	0.840	0.759	Doctor	0.521	0.512	0.965	0.669
	POS	Restaurant	0.735	0.697	0.815	0.751	Doctor	0.540	0.521	0.975	0.679
SAGE	Unigram	Restaurant	0.770	0.793	0.750	0.784	Doctor	0.520	0.547	0.705	0.616
	LIWC	Restaurant	0.742	0.728	0.749	0.738	Doctor	0.647	0.650	0.608	0.628
	POS	Restaurant	0.746	0.732	0.687	0.701	Doctor	0.634	0.623	0.682	0.651

Table 4: Classifier performance in cross-domain adaptation.

erties among restaurants and hotels. Among three types of features, Unigram still performs best. POS and LIWC features are also robust across domains.

In the doctor domain, we observe that models trained on Unigram features from the hotels domain do not generalize well to doctor reviews, and the performance is a little bit better than random guess with only 0.55 accuracy. For SVM, models trained on POS and LIWC features achieve even lower accuracy than Unigram. POS and LIWC features obtain around 0.5 precision and 1.0 recall, indicating that all doctor reviews are classified as deceptive by the classifier. One plausible explanation could be doctor reviews generally encode some type of positive-weighted (deceptive) features more than hotel reviews and these types of features dominate the decision making procedures, leading all reviews to be classified as deceptive.

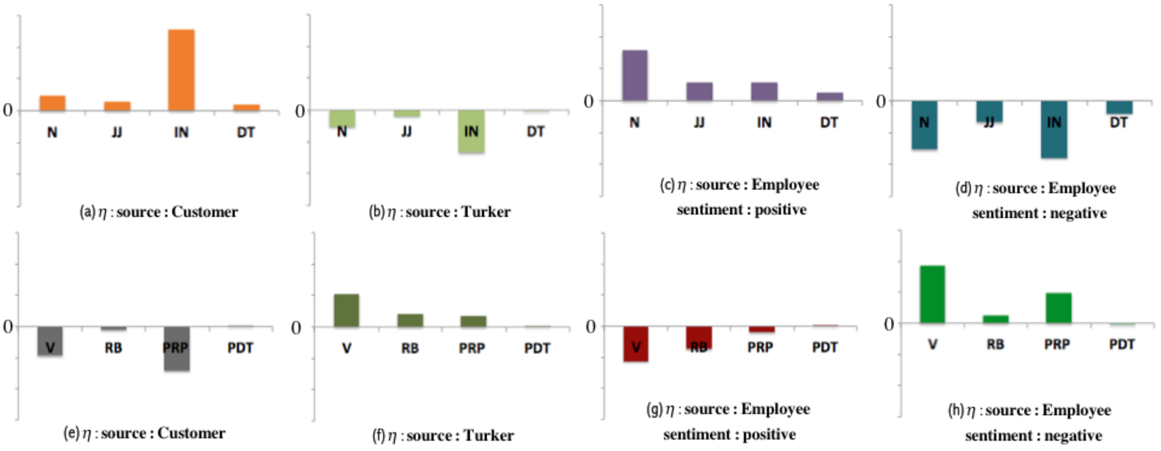
Tables 5 and 6 give the top weighted LIWC and POS features. We observe that many features are indeed shared among doctor and hotel domains. Notably, POS features are more robust than LIWC as more shared features are observed. As domain specific properties will be considered in the interaction part (η^{LIWC}_{domain} and η^{POS}_{domain}) of the additive model, SAGE achieve much better results than SVM, and is around 0.65 accurate in the cross-domain task.

LIWC (hotel)		LIWC (doctor)	
deceptive	truthful	deceptive	truthful
i	AllPct	Sixletters	present
family	number	past	AllPct
pronoun	hear	work	social
Sixletters	we	health	shehe
see	space	i	number
posemo	dash	friend	time
certain	human	posemo	we
leisure	exclusive	feel	you
future	past	perceptual	negemo
perceptual	home	leisure	Period
feel	otherpunct	insight	relativ
comma	negemo	comma	ingest
cause	dash	future	money

Table 5: Top weighted LIWC features for Turker vs Customer in Doctor and Hotel reviews. Blue denotes shared positive (deceptive) features and red denotes negative (truthful) features.

6 General Linguistic Cues of Deceptive Opinion Spam

In this section, we examine a number of general POS and LIWC features that may shed light on a general rule for identifying deceptive opinion



(a) η : source : Customer

(b) η : source : Turker

(c) η : source : Employee sentiment : positive

(d) η : source : Employee sentiment : negative

(e) η : source : Customer

(f) η : source : Turker

(g) η : source : Employee sentiment : positive

(h) η : source : Employee sentiment : negative

Figure 1: Visualization of the η for POS features: Horizontal axes correspond to the values η and are NORMALIZED from the log-frequency function.

POS (hotel)		POS (doctor)	
deceptive	truthful	deceptive	truthful
PRP$	CD	VBD	CD
PRP	RRB	NNP	VBZ
VB	LRB	VB	VBP
TO	CC	TO	FW
NNP	NNS	VBG	RRB
VBG	RP	PRP$	LRB
MD	VBN	JJS	RB
VBP	IN	JJ	LS
RB	EX	WRB	PDT
JJS	VBZ	PRP	VBN

Table 6: Top weighted POS features for Turker vs Customer in Doctor and Hotel reviews. Blue denotes shared positive (deceptive) features and red denotes negative (truthful) features.

spam. Our modified SAGE model provides us with a tailored tool for this analysis. Specifically, each feature f is associated with a background value m^f. For each facet A, η_A^f, presents the facet-specific preference value for feature f. Note that sentiments are separated into positive and negative dimensions, which is necessary because hotel employee authors wrote positive-sentiment reviews when reviewing their own hotels, and negative-sentiment reviews when reviewing their competitors' hotels.

6.1 POS features

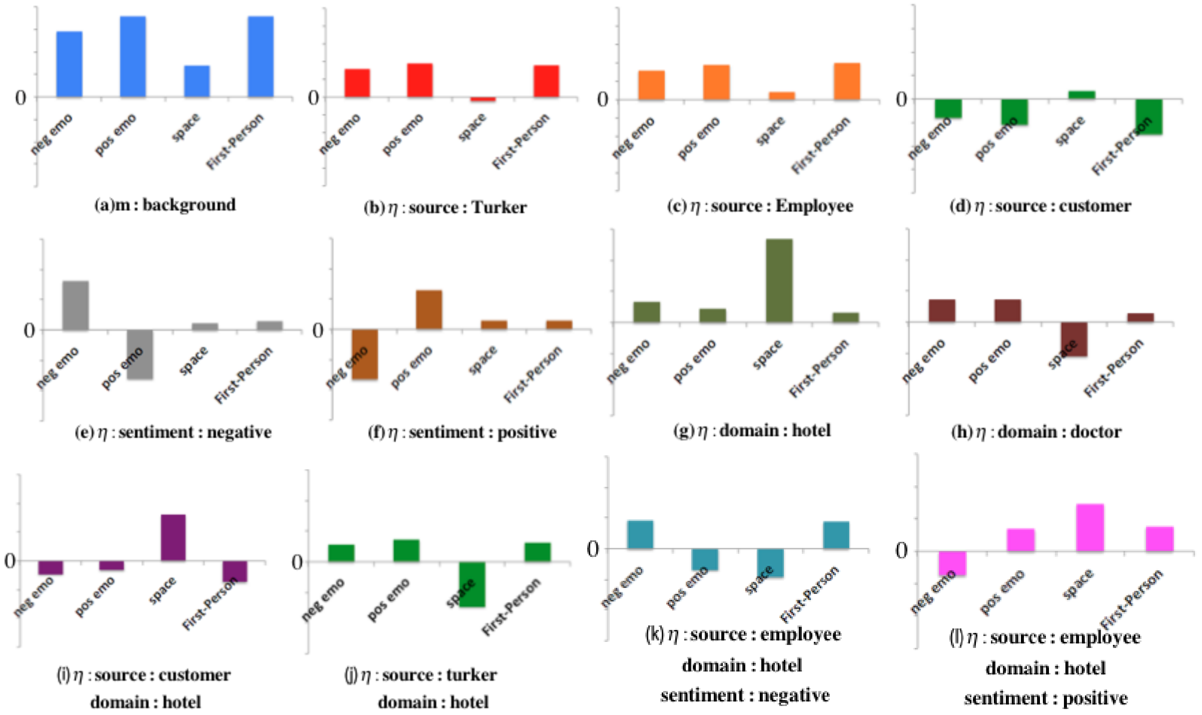
Early findings in the literature (Rayson et al., 2001; Buller and Burgoon, 1996; Biber et al., 1999) found that informative (truthful) writings typically consist of more nouns, adjectives, prepositions, determiners, and coordinating conjunctions, while imaginative (deceptive) writing consist of more verbs, adverbs, pronouns, and predeterminers (with a few exceptions). Our findings with POS features are largely in agreement with these findings when distinguishing between Turker and Customer reviews, but are violated in the Employee set.

We present the eight types of POS features in Figure 1, namely, N (Noun), JJ (Adjective), IN (Preposition or subordinating conjunction) and DT (Determiner), V (Verb), RB (Adverb), PRP (Pronouns, both personal and possessive) and PDT (Pre-Determiner).

From Figures 1(a)(b)(e)(f), we observe that with the exception of PDT, the word frequency of which is too small to draw a conclusion, *Turker* and *Customer* reviews exhibit linguistic patterns in agreement with previous findings in the literature, where truthful reviews (*Customer*) tend to include more N, JJ, IN and DT, while deceptive writings tend to encode more V, RB and PRP.

However, in the case of the *Employee-Positive* dataset, which is equally deceptive, most of these rules are violated. Notably, reviews from the *Employee-Positive* set did not encode fewer N, JJ and DT terms, as expected (see Figures 1(a)(c)). Instead, they encode even more N, JJ and DT vocabularies than truthful reviews from the *Customer* reviews. Also, fewer V and RB are found in *Employee-Positive* reviews compared with *Customer* reviews (see Figures 1(e)(g)).

One explanation for these observations is that informative (truthful) writing tends to be more introductory and descriptive, encoding more concrete details, when compared with imaginary writings. As domain experts possess considerable knowledge of their own offerings, they highlight



(a)m : background

(b) η : source : Turker

(c) η : source : Employee

(d) η : source : customer

(e) η : sentiment : negative

(f) η : sentiment : positive

(g) η : domain : hotel

(h) η : domain : doctor

(i) η : source : customer
domain : hotel

(j) η : source : turker
domain : hotel

(k) η : source : employee
domain : hotel
sentiment : negative

(l) η : source : employee
domain : hotel
sentiment : positive

Figure 2: Visualization of the η for LIWC features: Horizontal axes correspond to the values η and are normalized from the log-frequency function.

the details and their lies may be even more informative and descriptive than those generated by real customers! This explains why *Employee-Positive* contains more N, IN and DT. Meanwhile, as domain experts are engaged more in talking about the details, they inevitably overlook other information, possibly leading to fewer V and RB.

For *Employee-Positive* reviews, shown in Figures 1(d)(h), it turns out that domain experts do not compensate for their lack of prior experience when writing negative reviews for competitors' offerings, as we will see again with LIWC features in the next subsection.

6.2 LIWC features

We explore 3 LIWC categories (from left to right in subfigures of Figure 2): sentiment (*neg emo* and *pos emo*), spatial detail (*space*), and first-person singular pronouns (*first-person*).

Space: Note that spatial details are more specific in the *Hotel* and *Restaurant* domains, which is reflected in the high positive value of $\eta_{domain}^{Hotel,space}$ (see Figure 2(g)) and negative value of $\eta_{domain}^{Doctor,space}$ (see Figure 2(h)). It illustrates how domain-specific details can be predictive of deceptive text. Similarly predictive LIWC features are *home* for the *Hotel* domain, *ingest* for the *Restaurant* domain, and *health* and *body* for the *Doctor* domain.

In Figure 2(i)(j)(k)(l), we can easily see that both actual customers and domain experts encode more spatial details in their reviews (positive value of η), which is in agreement with our expectation. This further demonstrates that a lack of spatial details would not be a general cue for deception. Moreover, it appears that general domain expertise does not compensate for the lack of prior experience when writing deceptive negative reviews for competitors' hotels, as demonstrated by the lack of spatial details in the negative-sentiment reviews by employees shown in Figure 2(k).

Sentiment: According to our findings, the presence of sentiment is a general cue to deceptive opinion spam, as observed when comparing Figure 2(b) to Figure 2(c) and (d). Participants, both Employees and Turkers, tend to exaggerate sentiment, and include more sentiment-related vocabularies in their lies. In other words, positive deceptive reviews were generally more positive and negative deceptive reviews were more negative in sentiment when compared with the truthful reviews generated by actual customers. A similar pattern can also be observed when comparing Figure 2(i) to Figure 2(j).

First-Person Singular Pronouns: The literature also associates deception with decreased usage of first-person singular pronouns, an effect attributed to psychological distancing, whereby deceivers talk less about themselves due either to a lack of personal experience, or to detach themselves from the lie (Newman et al., 2003; Zhou et al., 2004; Buller et al., 1996; Knapp and Comaden, 1979). However, according to our findings, we find the opposite to hold. Increased first person singular is an apparent indicator of deception, when comparing Figure 2(b) to 2(c) and 2(e). We suspect that this relates to an effect observed in previous studies of deception, where liars inadvertently undermine their lies by overemphasizing aspects of their deception that they believe reflect credibility (Bond and DePaulo, 2006; DePaulo et al., 2003). One interpretation for this phenomenon would be that deceivers try to overemphasize their physical presence because they believe that this increases their credibility.

7 Conclusion and Discussion

In this work, we have developed a multi-domain large-scale dataset containing gold-standard deceptive opinion spam. It includes reviews of Hotels, Restaurants and Doctors, generated through crowdsourcing and domain experts. We study this data using SAGE, which enables us to make observations about the respects in which truthful and deceptive text differs. Our model includes several domain-independent features that shed light on these differences, which further allows us to formulate some general rules for recognizing deceptive opinion spam.

We also acknowledge several important caveats to this work. By soliciting fake reviews from participants, including crowd workers and domain experts, we have found that is possible to detect fake reviews with above-chance accuracy, and have used our models to explore several psychological theories of deception. However, it is still very difficult to estimate the practical impact of such methods, as it is very challenging to obtain gold-standard data in the real world. Moreover, by soliciting deceptive opinion spam in an artificial environment, we are endorsing the deception, which may influence the cues that we observe (Feeley and others, 1998; Frank and Ekman, 1997; Newman et al., 2003; Ott, 2013). Finally, it may be possible to train people to tell more convincing lies. Many of the characteristics regarding fake review generation might be overcome by well-trained fake review writers, which would results in opinion spam that is harder for detect. Future work may wish to consider some of these additional challenges.

8 Acknowledgement

We thank Wenjie Li and Xun Wang for useful discussions and suggestions. This work was supported in part by National Science Foundation Grant BCS-0904822, a DARPA Deft grant, as well as a gift from Google. We also thank the ACL reviewers for their helpful comments and advice.

References

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Randolph Quirk. 1999. *Longman grammar of spoken and written English*, volume 2. MIT Press.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Charles Bond and Bella DePaulo. 2006. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234.

David B Buller and Judee K Burgoon. 1996. Interpersonal deception theory. *Communication theory*, 6(3):203–242.

David B Buller, Judee K Burgoon, Aileen Buslig, and James Roiger. 1996. Testing interpersonal deception theory: The language of interpersonal deception. *Communication theory*, 6(3):268–289.

Paul-Alexandru Chirita, Jörg Diederich, and Wolfgang Nejdl. 2005. Mailrank: using ranking for spam detection. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 373–380. ACM.

Cone. 2011. 2011 Online Influence Trend Tracker. http://www.coneinc.com/negative-reviews-online-reverse-purchase-decisions, August.

Bella DePaulo, James Lindsay, Brian Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin*, 129(1):74.

Harris Drucker, Donghui Wu, and Vladimir Vapnik. 1999. Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on*, 10(5):1048–1054.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1041–1048.

Thomas Feeley. 1998. The behavioral correlates of sanctioned and unsanctioned deceptive communication. *Journal of Nonverbal Behavior*, 22(3):189–204.

Vanessa Feng and Graeme Hirst. 2013. Detecting deceptive opinions with profile compatibility. In *Proceedings of the 6th International Joint Conference on Natural Language Processing, Nagoya, Japan*, pages 14–18.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics.

Mark Frank and Paul Ekman. 1997. The ability to detect deceit generalizes across different types of high-stake lies. *Journal of personality and social psychology*, 72(6):1429.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 576–587. VLDB Endowment.

Trevor J Hastie and Robert J Tibshirani. 1990. *Generalized additive models*, volume 43. CRC Press.

Ipsos. 2012. Socialogue: Five Stars? Thumbs Up? A+ or Just Average? http://www.ipsos-na.com/news-polls/pressrelease.aspx?id=5929.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, pages 219–230. ACM.

Nitin Jindal, Bing Liu, and Ee-Peng Lim. 2010. Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1549–1552. ACM.

Thorsten Joachims. 1999. Making large scale svm learning practical.

Marcia K Johnson and Carol L Raye. 1981. Reality monitoring. *Psychological review*, 88(1):67.

Mark Knapp and Mark Comaden. 1979. Telling it like it isn't: A review of theory and research on deceptive communications. *Human Communication Research*, 5(3):270–285.

Jiwei Li, Claire Cardie, and Sujian Li. 2013a. Topicspam: a topic-model-based approach for spam detection. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguis-tics*.

Jiwei Li, Myle Ott, and Claire Cardie. 2013b. Identifying manipulated offerings on review portals. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 939–948. ACM.

Juan Martinez-Romo and Lourdes Araujo. 2009. Web spam identification through language model analysis. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, pages 21–28. ACM.

David Meyer. 2009. Fake reviews prompt belkin apology.

Claire Miller. 2009. Company settles case of reviews it faked. *New York Times*.

Arjun Mukherjee, Bing Liu, Junhui Wang, Natalie Glance, and Nitin Jindal. 2011. Detecting group review spam. In *Proceedings of the 20th international conference companion on World wide web*, pages 93–94. ACM.

Arjun Mukherjee, Bing Liu, and Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200. ACM.

Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013a. Spotting opinion spammers using behavioral footprints. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 632–640. ACM.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. 2013b. What yelp fake review filter might be doing. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, pages 83–92. ACM.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 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*, pages 309–319.

Myle Ott, Claire Cardie, and Jeff Hancock. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web*, pages 201–210. ACM.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Short Papers*, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Myle Ott. 2013. Computational lingustic models of deceptive opinion spam. *PHD, thesis*.

Sinno Pan and Qiang Yang. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

Tieyun Qian and Bing Liu. 2013. Identifying multiple userids of the same author. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Paul Rayson, Andrew Wilson, and Geoffrey Leech. 2001. Grammatical word class variation within the british national corpus sampler. *Language and Computers*, 36(1):295–306.

David Streitfeld. 2012. For 2 a star, an online retailer gets 5-star product reviews. *New York Times.*, 26.

Alexandra Topping. 2010. Historian orlando figes agrees to pay damages for fake reviews. *The Guardian.*, 16.

Guan Wang, Sihong Xie, Bing Liu, and Philip Yu. 2011. Review graph based online store review spammer detection. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 1242–1247. IEEE.

Guan Wang, Sihong Xie, Bing Liu, and Philip Yu. 2012. Identify online store review spammers via social review graph. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):61.

Guangyu Wu, Derek Greene, Barry Smyth, and Pádraig Cunningham. 2010. Distortion as a validation criterion in the identification of suspicious reviews. In *Proceedings of the First Workshop on Social Media Analytics*, pages 10–13. ACM.

Kyung-Hyan Yoo and Ulrike Gretzel. 2009. Comparison of deceptive and truthful travel reviews. In *Information and communication technologies in tourism 2009*, pages 37–47. Springer.

Lina Zhou, Judee K Burgoon, Douglas P Twitchell, Tiantian Qin, and Jay F Nunamaker Jr. 2004. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–166.