Finding Deceptive Opinion Spam by Any Stretch of the Imagination

Myle Ott **Yejin Choi** **Claire Cardie** **Jeffrey T. Hancock**
Department of Computer Science, Cornell University, Ithaca, NY 14853
Department of Communication, Cornell University, Ithaca, NY 14853
{myleott,ychoi,cardie}@cs.cornell.edu jth34@cornell.edu

Abstract

Consumers increasingly rate, review and research products online (Jansen, 2010; Litvin et al., 2008). Consequently, websites containing consumer reviews are becoming targets of *opinion spam*. While recent work has focused primarily on manually identifiable instances of opinion spam, in this work we study *deceptive opinion spam*—fictitious opinions that have been deliberately written to sound authentic. Integrating work from psychology and computational linguistics, we develop and compare three approaches to detecting deceptive opinion spam, and ultimately develop a classifier that is nearly 90% accurate on our *gold-standard* opinion spam dataset. Based on feature analysis of our learned models, we additionally make several theoretical contributions, including revealing a relationship between deceptive opinions and imaginative writing.

1 Introduction

With the ever-increasing popularity of review websites that feature user-generated opinions (e.g., TripAdvisor[1] and Yelp[2]), there comes an increasing potential for monetary gain through *opinion spam*—inappropriate or fraudulent reviews. Opinion spam can range from annoying self-promotion of an unrelated website or blog to deliberate review fraud, as in the recent case[3] of a Belkin employee who hired people to write positive reviews for an otherwise poorly reviewed product.[4]

While other kinds of spam have received considerable computational attention, regrettably there has been little work to date (see Section 2) on opinion spam detection. Furthermore, most previous work in the area has focused on the detection of DISRUPTIVE OPINION SPAM—uncontroversial instances of spam that are easily identified by a human reader, e.g., advertisements, questions, and other irrelevant or non-opinion text (Jindal and Liu, 2008). And while the presence of disruptive opinion spam is certainly a nuisance, the risk it poses to the user is minimal, since the user can always choose to ignore it.

We focus here on a potentially more insidious type of opinion spam: DECEPTIVE OPINION SPAM—fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader. For example, one of the following two hotel reviews is truthful and the other is *deceptive opinion spam*:

1. I have stayed at many hotels traveling for both business and pleasure and I can honestly stay that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to both business travellers and couples.
2. My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The area of the hotel is great, since I love to shop I couldn't ask for more!! We will definatly be

[1] http://tripadvisor.com
[2] http://yelp.com
[3] http://news.cnet.com/8301-1001_3-10145399-92.html
[4] It is also possible for opinion spam to be negative, potentially in order to sully the reputation of a competitor.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 309–319,
Portland, Oregon, June 19-24, 2011.

back to Chicago and we will for sure be back to the James Chicago.

Typically, these deceptive opinions are neither easily ignored nor even identifiable by a human reader;[5] consequently, there are few good sources of labeled data for this research. Indeed, in the absence of gold-standard data, related studies (see Section 2) have been forced to utilize ad hoc procedures for evaluation. In contrast, one contribution of the work presented here is the creation of the first large-scale, publicly available[6] dataset for deceptive opinion spam research, containing 400 truthful and 400 *gold-standard* deceptive reviews.

To obtain a deeper understanding of the nature of deceptive opinion spam, we explore the relative utility of three potentially complementary framings of our problem. Specifically, we view the task as: (a) a standard *text categorization* task, in which we use n-gram–based classifiers to label opinions as either deceptive or truthful (Joachims, 1998; Sebastiani, 2002); (b) an instance of *psycholinguistic deception detection*, in which we expect deceptive statements to exemplify the psychological effects of lying, such as increased negative emotion and psychological distancing (Hancock et al., 2008; Newman et al., 2003); and, (c) a problem of *genre identification*, in which we view deceptive and truthful writing as sub-genres of imaginative and informative writing, respectively (Biber et al., 1999; Rayson et al., 2001).

We compare the performance of each approach on our novel dataset. Particularly, we find that machine learning classifiers trained on features traditionally employed in (a) psychological studies of deception and (b) genre identification are both outperformed at statistically significant levels by n-gram–based text categorization techniques. Notably, a combined classifier with both n-gram and psychological deception features achieves nearly 90% cross-validated accuracy on this task. In contrast, we find deceptive opinion spam detection to be well beyond the capabilities of most human judges, who perform roughly at-chance—a finding that is consistent with decades of traditional deception detection research (Bond and DePaulo, 2006).

[5]The second example review is deceptive opinion spam.

[6]Available by request at: `http://www.cs.cornell.edu/~myleott/op_spam`

Additionally, we make several theoretical contributions based on an examination of the feature weights learned by our machine learning classifiers. Specifically, we shed light on an ongoing debate in the deception literature regarding the importance of considering the context and motivation of a deception, rather than simply identifying a universal set of deception cues. We also present findings that are consistent with recent work highlighting the difficulties that liars have encoding spatial information (Vrij et al., 2009). Lastly, our study of deceptive opinion spam detection as a genre identification problem reveals relationships between deceptive opinions and imaginative writing, and between truthful opinions and informative writing.

The rest of this paper is organized as follows: in Section 2, we summarize related work; in Section 3, we explain our methodology for gathering data and evaluate human performance; in Section 4, we describe the features and classifiers employed by our three automated detection approaches; in Section 5, we present and discuss experimental results; finally, conclusions and directions for future work are given in Section 6.

2 Related Work

Spam has historically been studied in the contexts of e-mail (Drucker et al., 2002), and the Web (Gyöngyi et al., 2004; Ntoulas et al., 2006). Recently, researchers have began to look at *opinion spam* as well (Jindal and Liu, 2008; Wu et al., 2010; Yoo and Gretzel, 2009).

Jindal and Liu (2008) find that opinion spam is both widespread and different in nature from either e-mail or Web spam. Using product review data, and in the absence of gold-standard deceptive opinions, they train models using features based on the review text, reviewer, and product, to distinguish between *duplicate* opinions[7] (considered deceptive spam) and *non-duplicate* opinions (considered truthful). Wu et al. (2010) propose an alternative strategy for detecting deceptive opinion spam in the absence

[7]Duplicate (or near-duplicate) opinions are opinions that appear more than once in the corpus with the same (or similar) text. While these opinions are likely to be deceptive, they are unlikely to be representative of deceptive opinion spam in general. Moreover, they are potentially detectable via off-the-shelf plagiarism detection software.

of gold-standard data, based on the distortion of popularity rankings. Both of these heuristic evaluation approaches are unnecessary in our work, since we compare *gold-standard* deceptive and truthful opinions.

Yoo and Gretzel (2009) gather 40 truthful and 42 deceptive hotel reviews and, using a standard statistical test, manually compare the psychologically relevant linguistic differences between them. In contrast, we create a much larger dataset of 800 opinions that we use to develop and evaluate *automated* deception classifiers.

Research has also been conducted on the related task of *psycholinguistic deception detection*. Newman et al. (2003), and later Mihalcea and Strapparava (2009), ask participants to give both their true and untrue views on personal issues (e.g., their stance on the death penalty). Zhou et al. (2004; 2008) consider computer-mediated deception in role-playing games designed to be played over instant messaging and e-mail. However, while these studies compare n-gram–based deception classifiers to a random guess baseline of 50%, we additionally evaluate and compare two other computational approaches (described in Section 4), as well as the performance of human judges (described in Section 3.3).

Lastly, automatic approaches to determining *review quality* have been studied—directly (Weimer et al., 2007), and in the contexts of helpfulness (Danescu-Niculescu-Mizil et al., 2009; Kim et al., 2006; O'Mahony and Smyth, 2009) and credibility (Weerkamp and De Rijke, 2008). Unfortunately, most measures of quality employed in those works are based exclusively on human judgments, which we find in Section 3 to be poorly calibrated to detecting deceptive opinion spam.

3 Dataset Construction and Human Performance

While truthful opinions are ubiquitous online, deceptive opinions are difficult to obtain without resorting to heuristic methods (Jindal and Liu, 2008; Wu et al., 2010). In this section, we report our efforts to gather (and validate with human judgments) the first publicly available opinion spam dataset with *gold-standard* deceptive opinions.

Following the work of Yoo and Gretzel (2009), we compare truthful and deceptive **positive** reviews for hotels found on TripAdvisor. Specifically, we mine all 5-star truthful reviews from the 20 most popular hotels on TripAdvisor[8] in the Chicago area.[9] Deceptive opinions are gathered for those same 20 hotels using Amazon Mechanical Turk[10] (AMT). Below, we provide details of the collection methodologies for deceptive (Section 3.1) and truthful opinions (Section 3.2). Ultimately, we collect 20 truthful and 20 deceptive opinions for each of the 20 chosen hotels (800 opinions total).

3.1 Deceptive opinions via Mechanical Turk

Crowdsourcing services such as AMT have made large-scale data annotation and collection efforts financially affordable by granting anyone with basic programming skills access to a marketplace of anonymous online workers (known as *Turkers*) willing to complete small tasks.

To solicit gold-standard **deceptive** opinion spam using AMT, we create a pool of 400 *Human Intelligence Tasks* (HITs) and allocate them evenly across our 20 chosen hotels. To ensure that opinions are written by unique authors, we allow only a single submission per Turker. We also restrict our task to Turkers who are located in the United States, and who maintain an approval rating of at least 90%. Turkers are allowed a maximum of 30 minutes to work on the HIT, and are paid one US dollar for an accepted submission.

Each HIT presents the Turker with the name and website of a hotel. The HIT instructions ask the Turker to assume that they work for the hotel's marketing department, and to pretend that their boss wants them to write a fake review (as if they were a customer) to be posted on a travel review website; additionally, the review needs to sound realistic and portray the hotel in a positive light. A disclaimer

[8] TripAdvisor utilizes a proprietary ranking system to assess hotel popularity. We chose the 20 hotels with the greatest number of reviews, irrespective of the TripAdvisor ranking.

[9] It has been hypothesized that popular offerings are less likely to become targets of deceptive opinion spam, since the relative impact of the spam in such cases is small (Jindal and Liu, 2008; Lim et al., 2010). By considering only the most popular hotels, we hope to minimize the risk of mining opinion spam and labeling it as truthful.

[10] `http://mturk.com`

Time spent t (minutes)	
All submissions	$count$: 400 t_{min}: 0.08, t_{max}: 29.78 $\bar{t}$: 8.06, s: 6.32
Length ℓ (words)	
All submissions	ℓ_{min}: 25, ℓ_{max}: 425 $\bar{\ell}$: 115.75, s: 61.30
Time spent $t < 1$	$count$: 47 ℓ_{min}: 39, ℓ_{max}: 407 $\bar{\ell}$: 113.94, s: 66.24
Time spent $t \geq 1$	$count$: 353 ℓ_{min}: 25, ℓ_{max}: 425 $\bar{\ell}$: 115.99, s: 60.71

Table 1: Descriptive statistics for 400 deceptive opinion spam submissions gathered using AMT. s corresponds to the sample standard deviation.

indicates that any submission found to be of insufficient quality (e.g., written for the wrong hotel, unintelligible, unreasonably short,[11] plagiarized,[12] etc.) will be rejected.

It took approximately 14 days to collect 400 satisfactory deceptive opinions. Descriptive statistics appear in Table 1. Submissions vary quite dramatically both in length, and time spent on the task. Particularly, nearly 12% of the submissions were completed in *under one minute*. Surprisingly, an independent two-tailed t-test between the mean length of these submissions ($\bar{\ell}_{t<1}$) and the other submissions ($\bar{\ell}_{t\geq 1}$) reveals no significant difference ($p = 0.83$). We suspect that these "*quick*" users may have started working prior to having formally accepted the HIT, presumably to circumvent the imposed time limit. Indeed, the quickest submission took just 5 seconds and contained 114 words.

3.2 Truthful opinions from TripAdvisor

For truthful opinions, we mine all 6,977 reviews from the 20 most popular Chicago hotels on TripAdvisor. From these we eliminate:

- 3,130 non-5-star reviews;
- 41 non-English reviews;[13]
- 75 reviews with fewer than 150 characters since, by construction, deceptive opinions are at least 150 characters long (see footnote 11 in Section 3.1);
- 1,607 reviews written by *first-time authors*—new users who have not previously posted an opinion on TripAdvisor—since these opinions are more likely to contain opinion spam, which would reduce the integrity of our truthful review data (Wu et al., 2010).

Finally, we balance the number of truthful and deceptive opinions by selecting 400 of the remaining 2,124 truthful reviews, such that the document lengths of the selected truthful reviews are similarly distributed to those of the deceptive reviews. Work by Serrano et al. (2009) suggests that a *log-normal* distribution is appropriate for modeling document lengths. Thus, for each of the 20 chosen hotels, we select 20 truthful reviews from a log-normal (left-truncated at 150 characters) distribution fit to the lengths of the deceptive reviews.[14] Combined with the 400 deceptive reviews gathered in Section 3.1 this yields our final dataset of 800 reviews.

3.3 Human performance

Assessing human deception detection performance is important for several reasons. First, there are few other baselines for our classification task; indeed, related studies (Jindal and Liu, 2008; Mihalcea and Strapparava, 2009) have only considered a random guess baseline. Second, assessing human performance is necessary to validate the deceptive opinions gathered in Section 3.1. If human performance is low, then our deceptive opinions are convincing, and therefore, deserving of further attention.

Our initial approach to assessing human performance on this task was with Mechanical Turk. Unfortunately, we found that some Turkers selected among the choices seemingly at random, presumably to maximize their hourly earnings by obviating the need to read the review. While a similar effect has been observed previously (Akkaya et al., 2010), there remains no universal solution.

Instead, we solicit the help of three volunteer undergraduate university students to make judgments on a subset of our data. This balanced subset, corresponding to the first fold of our cross-validation

[11] A submission is considered unreasonably short if it contains fewer than 150 characters.

[12] Submissions are individually checked for plagiarism at `http://plagiarisma.net`.

[13] Language is determined using `http://tagthe.net`.

[14] We use the R package GAMLSS (Rigby and Stasinopoulos, 2005) to fit the left-truncated log-normal distribution.

			TRUTHFUL			DECEPTIVE		
		Accuracy	**P**	**R**	**F**	**P**	**R**	**F**
HUMAN	JUDGE 1	**61.9%**	57.9	87.5	**69.7**	74.4	36.3	48.7
	JUDGE 2	56.9%	53.9	**95.0**	68.8	**78.9**	18.8	30.3
	JUDGE 3	53.1%	52.3	70.0	59.9	54.7	36.3	43.6
META	MAJORITY	58.1%	54.8	92.5	68.8	76.0	23.8	36.2
	SKEPTIC	60.6%	**60.8**	60.0	60.4	60.5	**61.3**	**60.9**

Table 2: Performance of three human judges and two meta-judges on a subset of 160 opinions, corresponding to the first fold of our cross-validation experiments in Section 5. Boldface indicates the largest value for each column.

experiments described in Section 5, contains all 40 reviews from each of four randomly chosen hotels. Unlike the Turkers, our student volunteers are not offered a monetary reward. Consequently, we consider their judgements to be more honest than those obtained via AMT.

Additionally, to test the extent to which the individual human judges are biased, we evaluate the performance of two virtual meta-judges. Specifically, the MAJORITY meta-judge predicts "*deceptive*" when at least two out of three human judges believe the review to be deceptive, and the SKEPTIC meta-judge predicts "*deceptive*" when *any* human judge believes the review to be deceptive.

Human and meta-judge performance is given in Table 2. It is clear from the results that human judges are not particularly effective at this task. Indeed, a two-tailed binomial test fails to reject the null hypothesis that JUDGE 2 and JUDGE 3 perform at-chance ($p = 0.003, 0.10, 0.48$ for the three judges, respectively). Furthermore, all three judges suffer from *truth-bias* (Vrij, 2008), a common finding in deception detection research in which human judges are more likely to classify an opinion as truthful than deceptive. In fact, JUDGE 2 classified fewer than 12% of the opinions as deceptive! Interestingly, this bias is effectively smoothed by the SKEPTIC meta-judge, which produces nearly perfectly class-balanced predictions. A subsequent reevaluation of human performance on this task suggests that the truth-bias can be reduced if judges are given the class-proportions in advance, although such prior knowledge is unrealistic; and ultimately, performance remains similar to that of Table 2.

Inter-annotator agreement among the three judges, computed using Fleiss' kappa, is 0.11. While there is no precise rule for interpreting kappa scores, Landis and Koch (1977) suggest that scores in the range (0.00, 0.20] correspond to "*slight agreement*" between annotators. The largest pairwise Cohen's kappa is 0.12, between JUDGE 2 and JUDGE 3—a value far below generally accepted pairwise agreement levels. We suspect that agreement among our human judges is so low *precisely because* humans are poor judges of deception (Vrij, 2008), and therefore they perform nearly at-chance respective to one another.

4 Automated Approaches to Deceptive Opinion Spam Detection

We consider three automated approaches to detecting deceptive opinion spam, each of which utilizes classifiers (described in Section 4.4) trained on the dataset of Section 3. The features employed by each strategy are outlined here.

4.1 Genre identification

Work in computational linguistics has shown that the frequency distribution of *part-of-speech* (POS) tags in a text is often dependent on the genre of the text (Biber et al., 1999; Rayson et al., 2001). In our genre identification approach to deceptive opinion spam detection, we test if such a relationship exists for truthful and deceptive reviews by constructing, for each review, features based on the frequencies of each POS tag.[15] These features are also intended to provide a good baseline with which to compare our other automated approaches.

4.2 Psycholinguistic deception detection

The *Linguistic Inquiry and Word Count* (LIWC) software (Pennebaker et al., 2007) is a popular automated text analysis tool used widely in the social sciences. It has been used to detect personality

[15] We use the Stanford Parser (Klein and Manning, 2003) to obtain the relative POS frequencies.

traits (Mairesse et al., 2007), to study tutoring dynamics (Cade et al., 2010), and, most relevantly, to analyze deception (Hancock et al., 2008; Mihalcea and Strapparava, 2009; Vrij et al., 2007).

While LIWC does not include a text classifier, we can create one with features derived from the LIWC output. In particular, LIWC counts and groups the number of instances of nearly 4,500 keywords into 80 psychologically meaningful dimensions. We construct one feature for each of the 80 LIWC dimensions, which can be summarized broadly under the following four categories:

1. Linguistic processes: Functional aspects of text (e.g., the average number of words per sentence, the rate of misspelling, swearing, etc.)
2. Psychological processes: Includes all social, emotional, cognitive, perceptual and biological processes, as well as anything related to time or space.
3. Personal concerns: Any references to work, leisure, money, religion, etc.
4. Spoken categories: Primarily filler and agreement words.

While other features have been considered in past deception detection work, notably those of Zhou et al. (2004), early experiments found LIWC features to perform best. Indeed, the LIWC2007 software used in our experiments subsumes most of the features introduced in other work. Thus, we focus our psycholinguistic approach to deception detection on LIWC-based features.

4.3 Text categorization

In contrast to the other strategies just discussed, our text categorization approach to deception detection allows us to model both content and context with n-gram features. Specifically, we consider the following three n-gram feature sets, with the corresponding features lowercased and unstemmed: UNIGRAMS, BIGRAMS$^+$, TRIGRAMS$^+$, where the superscript $^+$ indicates that the feature set subsumes the preceding feature set.

4.4 Classifiers

Features from the three approaches just introduced are used to train Naïve Bayes and Support Vector Machine classifiers, both of which have performed well in related work (Jindal and Liu, 2008; Mihalcea and Strapparava, 2009; Zhou et al., 2008).

For a document $\vec{x}$, with label y, the *Naïve Bayes* (NB) classifier gives us the following decision rule:

$$\hat{y} = \arg\max_c \Pr(y = c) \cdot \Pr(\vec{x} \mid y = c) \quad (1)$$

When the class prior is *uniform*, for example when the classes are balanced (as in our case), (1) can be simplified to the maximum likelihood classifier (Peng and Schuurmans, 2003):

$$\hat{y} = \arg\max_c \Pr(\vec{x} \mid y = c) \quad (2)$$

Under (2), both the NB classifier used by Mihalcea and Strapparava (2009) and the language model classifier used by Zhou et al. (2008) are equivalent. Thus, following Zhou et al. (2008), we use the SRI Language Modeling Toolkit (Stolcke, 2002) to estimate individual language models, $\Pr(\vec{x} \mid y = c)$, for truthful and deceptive opinions. We consider all three n-gram feature sets, namely UNIGRAMS, BIGRAMS$^+$, and TRIGRAMS$^+$, with corresponding language models smoothed using the interpolated Kneser-Ney method (Chen and Goodman, 1996).

We also train *Support Vector Machine* (SVM) classifiers, which find a high-dimensional separating hyperplane between two groups of data. To simplify feature analysis in Section 5, we restrict our evaluation to *linear* SVMs, which learn a weight vector $\vec{w}$ and bias term b, such that a document $\vec{x}$ can be classified by:

$$\hat{y} = sign(\vec{w} \cdot \vec{x} + b) \quad (3)$$

We use SVM^{light} (Joachims, 1999) to train our linear SVM models on all three approaches and feature sets described above, namely POS, LIWC, UNIGRAMS, BIGRAMS$^+$, and TRIGRAMS$^+$. We also evaluate every combination of these features, but for brevity include only LIWC+BIGRAMS$^+$, which performs best. Following standard practice, document vectors are normalized to unit-length. For LIWC+BIGRAMS$^+$, we unit-length normalize LIWC and BIGRAMS$^+$ features individually before combining them.

			TRUTHFUL			DECEPTIVE		
Approach	**Features**	**Accuracy**	**P**	**R**	**F**	**P**	**R**	**F**
GENRE IDENTIFICATION	POS_{SVM}	73.0%	75.3	68.5	71.7	71.1	77.5	74.2
PSYCHOLINGUISTIC DECEPTION DETECTION	LIWC_{SVM}	76.8%	77.2	76.0	76.6	76.4	77.5	76.9
TEXT CATEGORIZATION	$\text{UNIGRAMS}_{\text{SVM}}$	88.4%	89.9	86.5	88.2	87.0	90.3	88.6
	$\text{BIGRAMS}^{+}_{\text{SVM}}$	89.6%	90.1	89.0	89.6	89.1	90.3	89.7
	$\text{LIWC+BIGRAMS}^{+}_{\text{SVM}}$	**89.8%**	89.8	**89.8**	**89.8**	**89.8**	89.8	**89.8**
	$\text{TRIGRAMS}^{+}_{\text{SVM}}$	89.0%	89.0	89.0	89.0	89.0	89.0	89.0
	$\text{UNIGRAMS}_{\text{NB}}$	88.4%	**92.5**	83.5	87.8	85.0	**93.3**	88.9
	$\text{BIGRAMS}^{+}_{\text{NB}}$	88.9%	89.8	87.8	88.7	88.0	90.0	89.0
	$\text{TRIGRAMS}^{+}_{\text{NB}}$	87.6%	87.7	87.5	87.6	87.5	87.8	87.6
HUMAN / META	JUDGE 1	**61.9%**	57.9	87.5	**69.7**	74.4	36.3	48.7
	JUDGE 2	56.9%	53.9	**95.0**	68.8	**78.9**	18.8	30.3
	SKEPTIC	60.6%	**60.8**	60.0	60.4	60.5	**61.3**	**60.9**

Table 3: Automated classifier performance for three approaches based on nested 5-fold cross-validation experiments. Reported precision, recall and F-score are computed using a micro-average, i.e., from the *aggregate* true positive, false positive and false negative rates, as suggested by Forman and Scholz (2009). Human performance is repeated here for JUDGE 1, JUDGE 2 and the SKEPTIC meta-judge, although they cannot be directly compared since the 160-opinion subset on which they are assessed only corresponds to the first cross-validation fold.

5 Results and Discussion

The deception detection strategies described in Section 4 are evaluated using a 5-fold *nested* cross-validation (CV) procedure (Quadrianto et al., 2009), where model parameters are selected for each test fold based on *standard* CV experiments on the training folds. Folds are selected so that each contains *all* reviews from four hotels; thus, learned models are always evaluated on reviews from unseen hotels.

Results appear in Table 3. We observe that automated classifiers outperform human judges for every metric, except truthful recall where JUDGE 2 performs best.[16] However, this is expected given that untrained humans often focus on unreliable cues to deception (Vrij, 2008). For example, one study examining deception in online dating found that humans perform at-chance detecting deceptive profiles because they rely on text-based cues that are unrelated to deception, such as second-person pronouns (Toma and Hancock, In Press).

Among the automated classifiers, baseline performance is given by the simple genre identification approach (POS_{SVM}) proposed in Section 4.1. Surprisingly, we find that even this simple automated classifier outperforms most human judges (one-tailed sign test $p = 0.06, 0.01, 0.001$ for the three judges, respectively, on the first fold). This result is best explained by theories of reality monitoring (Johnson and Raye, 1981), which suggest that truthful and deceptive opinions might be classified into informative and imaginative genres, respectively. Work by Rayson et al. (2001) has found strong distributional differences between informative and imaginative writing, namely that the former typically consists of more nouns, adjectives, prepositions, determiners, and coordinating conjunctions, while the latter consists of more verbs,[17] adverbs,[18] pronouns, and pre-determiners. Indeed, we find that the weights learned by POS_{SVM} (found in Table 4) are largely in agreement with these findings, notably except for adjective and adverb *superlatives*, the latter of which was found to be an exception by Rayson et al. (2001). However, that deceptive opinions contain more superlatives is not unexpected, since deceptive writing (but not necessarily imaginative writing in general) often contains exaggerated language (Buller and Burgoon, 1996; Hancock et al., 2008).

Both remaining automated approaches to detecting deceptive opinion spam outperform the simple

[16] As mentioned in Section 3.3, JUDGE 2 classified fewer than 12% of opinions as deceptive. While achieving 95% truthful recall, this judge's corresponding precision was not significantly better than chance (two-tailed binomial $p = 0.4$).

[17] *Past participle* verbs were an exception.

[18] *Superlative* adverbs were an exception.

TRUTHFUL/INFORMATIVE			DECEPTIVE/IMAGINATIVE		
Category	**Variant**	**Weight**	**Category**	**Variant**	**Weight**
NOUNS	Singular	0.008	VERBS	Base	-0.057
	Plural	0.002		Past tense	**0.041**
	Proper, singular	**-0.041**		Present participle	-0.089
	Proper, plural	0.091		Singular, present	-0.031
ADJECTIVES	General	0.002		Third person singular, present	**0.026**
	Comparative	0.058			
	Superlative	**-0.164**		Modal	-0.063
PREPOSITIONS	General	0.064	ADVERBS	General	**0.001**
DETERMINERS	General	0.009		Comparative	-0.035
COORD. CONJ.	General	0.094	PRONOUNS	Personal	-0.098
VERBS	Past participle	0.053		Possessive	-0.303
ADVERBS	Superlative	**-0.094**	PRE-DETERMINERS	General	**0.017**

Table 4: Average feature weights learned by POS_{SVM}. Based on work by Rayson et al. (2001), we expect weights on the left to be positive (predictive of *truthful* opinions), and weights on the right to be negative (predictive of *deceptive* opinions). Boldface entries are at odds with these expectations. We report average feature weights of *unit-normalized* weight vectors, rather than *raw* weights vectors, to account for potential differences in magnitude between the folds.

genre identification baseline just discussed. Specifically, the psycholinguistic approach (LIWC_{SVM}) proposed in Section 4.2 performs 3.8% more accurately (one-tailed sign test $p = 0.02$), and the standard text categorization approach proposed in Section 4.3 performs between 14.6% and 16.6% more accurately. However, best performance overall is achieved by combining features from these two approaches. Particularly, the combined model $\text{LIWC+BIGRAMS}^{+}_{\text{SVM}}$ is 89.8% accurate at detecting deceptive opinion spam.[19]

Surprisingly, models trained only on UNIGRAMS—the simplest n-gram feature set—outperform all non–text-categorization approaches, and models trained on BIGRAMS^{+} perform *even better* (one-tailed sign test $p = 0.07$). This suggests that a universal set of keyword-based deception cues (e.g., LIWC) is not the best approach to detecting deception, and a context-sensitive approach (e.g., BIGRAMS^{+}) might be necessary to achieve state-of-the-art deception detection performance.

To better understand the models learned by these automated approaches, we report in Table 5 the top 15 highest weighted features for each class (*truthful* and *deceptive*) as learned by $\text{LIWC+BIGRAMS}^{+}_{\text{SVM}}$ and LIWC_{SVM}. In agreement with theories of reality monitoring (Johnson and Raye, 1981), we observe that truthful opinions tend to include more sensorial and concrete language than deceptive opinions; in particular, truthful opinions are more specific about spatial configurations (e.g., small, bathroom, on, location). This finding is also supported by recent work by Vrij et al. (2009) suggesting that liars have considerable difficultly encoding spatial information into their lies. Accordingly, we observe an increased focus in deceptive opinions on aspects external to the hotel being reviewed (e.g., husband, business,

[19]The result is not significantly better than $\text{BIGRAMS}^{+}_{\text{SVM}}$.

$\text{LIWC+BIGRAMS}^{+}_{\text{SVM}}$		LIWC_{SVM}	
TRUTHFUL	DECEPTIVE	TRUTHFUL	DECEPTIVE
-	chicago	hear	i
...	my	number	family
on	hotel	allpunct	perspron
location	,_and	negemo	see
)	luxury	dash	pronoun
$\text{allpunct}_{\text{LIWC}}$	experience	exclusive	leisure
floor	hilton	we	exclampunct
(	business	sexual	sixletters
the_hotel	vacation	period	posemo
bathroom	i	otherpunct	comma
small	spa	space	cause
helpful	looking	human	auxverb
$	while	past	future
hotel_.	husband	inhibition	perceptual
other	my_husband	assent	feel

Table 5: Top 15 highest weighted truthful and deceptive features learned by $\text{LIWC+BIGRAMS}^{+}_{\text{SVM}}$ and LIWC_{SVM}. Ambiguous features are subscripted to indicate the source of the feature. LIWC features correspond to groups of keywords as explained in Section 4.2; more details about LIWC and the LIWC categories are available at `http://liwc.net`.

vacation).

We also acknowledge several findings that, on the surface, are in contrast to previous psycholinguistic studies of deception (Hancock et al., 2008; Newman et al., 2003). For instance, while deception is often associated with negative emotion terms, our deceptive reviews have more positive and fewer negative emotion terms. This pattern makes sense when one considers the goal of our deceivers, namely to create a positive review (Buller and Burgoon, 1996).

Deception has also previously been associated with decreased usage of first person singular, an effect attributed to psychological distancing (Newman et al., 2003). In contrast, we find increased first person singular to be among the largest indicators of deception, which we speculate is due to our deceivers attempting to enhance the credibility of their reviews by emphasizing their own presence in the review. Additional work is required, but these findings further suggest the importance of moving beyond a universal set of deceptive language features (e.g., LIWC) by considering both the contextual (e.g., BIGRAMS$^+$) and motivational parameters underlying a deception as well.

6 Conclusion and Future Work

In this work we have developed the first large-scale dataset containing *gold-standard* deceptive opinion spam. With it, we have shown that the detection of deceptive opinion spam is well beyond the capabilities of human judges, most of whom perform roughly at-chance. Accordingly, we have introduced three *automated* approaches to deceptive opinion spam detection, based on insights coming from research in computational linguistics and psychology. We find that while standard n-gram–based text categorization is the best individual detection approach, a *combination* approach using psycholinguistically-motivated features and n-gram features can perform slightly better.

Finally, we have made several theoretical contributions. Specifically, our findings suggest the importance of considering both the context (e.g., BIGRAMS$^+$) and motivations underlying a deception, rather than strictly adhering to a universal set of deception cues (e.g., LIWC). We have also presented results based on the feature weights learned by our classifiers that illustrate the difficulties faced by liars in encoding spatial information. Lastly, we have discovered a plausible relationship between deceptive opinion spam and imaginative writing, based on POS distributional similarities.

Possible directions for future work include an extended evaluation of the methods proposed in this work to both negative opinions, as well as opinions coming from other domains. Many additional approaches to detecting deceptive opinion spam are also possible, and a focus on approaches with high deceptive precision might be useful for production environments.

Acknowledgments

This work was supported in part by National Science Foundation Grants BCS-0624277, BCS-0904822, HSD-0624267, IIS-0968450, and NSCC-0904822, as well as a gift from Google, and the Jack Kent Cooke Foundation. We also thank, alphabetically, Rachel Boochever, Cristian Danescu-Niculescu-Mizil, Alicia Granstein, Ulrike Gretzel, Danielle Kirshenblat, Lillian Lee, Bin Lu, Jack Newton, Melissa Sackler, Mark Thomas, and Angie Yoo, as well as members of the Cornell NLP seminar group and the ACL reviewers for their insightful comments, suggestions and advice on various aspects of this work.

References

C. Akkaya, A. Conrad, J. Wiebe, and R. Mihalcea. 2010. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazons Mechanical Turk, Los Angeles*, pages 195–203.

D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, and R. Quirk. 1999. *Longman grammar of spoken and written English*, volume 2. MIT Press.

C.F. Bond and B.M. DePaulo. 2006. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214.

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

W.L. Cade, B.A. Lehman, and A. Olney. 2010. An exploration of off topic conversation. In *Human Language Technologies: The 2010 Annual Conference of*

the North American Chapter of the Association for Computational Linguistics, pages 669–672. Association for Computational Linguistics.

S.F. Chen and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics.

C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. 2009. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, pages 141–150. ACM.

H. Drucker, D. Wu, and V.N. Vapnik. 2002. Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on*, 10(5):1048–1054.

G. Forman and M. Scholz. 2009. Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement. *ACM SIGKDD Explorations*, 12(1):49–57.

Z. Gyöngyi, H. Garcia-Molina, and J. 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.

J.T. Hancock, L.E. Curry, S. Goorha, and M. Woodworth. 2008. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23.

J. Jansen. 2010. Online product research. *Pew Internet & American Life Project Report*.

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

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142.

T. Joachims. 1999. Making large-scale support vector machine learning practical. In *Advances in kernel methods*, page 184. MIT Press.

M.K. Johnson and C.L. Raye. 1981. Reality monitoring. *Psychological Review*, 88(1):67–85.

S.M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430. Association for Computational Linguistics.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

E.P. Lim, V.A. Nguyen, N. Jindal, B. Liu, and H.W. 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.

S.W. Litvin, R.E. Goldsmith, and B. Pan. 2008. Electronic word-of-mouth in hospitality and tourism management. *Tourism management*, 29(3):458–468.

F. Mairesse, M.A. Walker, M.R. Mehl, and R.K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500.

R. Mihalcea and C. Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Association for Computational Linguistics.

M.L. Newman, J.W. Pennebaker, D.S. Berry, and J.M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665.

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

M.P. O'Mahony and B. Smyth. 2009. Learning to recommend helpful hotel reviews. In *Proceedings of the third ACM conference on Recommender systems*, pages 305–308. ACM.

F. Peng and D. Schuurmans. 2003. Combining naive Bayes and n-gram language models for text classification. *Advances in Information Retrieval*, pages 547–547.

J.W. Pennebaker, C.K. Chung, M. Ireland, A. Gonzales, and R.J. Booth. 2007. The development and psychometric properties of LIWC2007. *Austin, TX, LIWC. Net*.

N. Quadrianto, A.J. Smola, T.S. Caetano, and Q.V. Le. 2009. Estimating labels from label proportions. *The Journal of Machine Learning Research*, 10:2349–2374.

P. Rayson, A. Wilson, and G. Leech. 2001. Grammatical word class variation within the British National Corpus sampler. *Language and Computers*, 36(1):295–306.

R.A. Rigby and D.M. Stasinopoulos. 2005. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

M.Á. Serrano, A. Flammini, and F. Menczer. 2009. Modeling statistical properties of written text. *PloS one*, 4(4):5372.

A. Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 3, pages 901–904. Citeseer.

C. Toma and J.T. Hancock. In Press. What Lies Beneath: The Linguistic Traces of Deception in Online Dating Profiles. *Journal of Communication*.

A. Vrij, S. Mann, S. Kristen, and R.P. Fisher. 2007. Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior*, 31(5):499–518.

A. Vrij, S. Leal, P.A. Granhag, S. Mann, R.P. Fisher, J. Hillman, and K. Sperry. 2009. Outsmarting the liars: The benefit of asking unanticipated questions. *Law and human behavior*, 33(2):159–166.

A. Vrij. 2008. *Detecting lies and deceit: Pitfalls and opportunities*. Wiley-Interscience.

W. Weerkamp and M. De Rijke. 2008. Credibility improves topical blog post retrieval. *ACL-08: HLT*, pages 923–931.

M. Weimer, I. Gurevych, and M. Mühlhäuser. 2007. Automatically assessing the post quality in online discussions on software. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 125–128. Association for Computational Linguistics.

G. Wu, D. Greene, B. Smyth, and P. Cunningham. 2010. Distortion as a validation criterion in the identification of suspicious reviews. Technical report, UCD-CSI-2010-04, University College Dublin.

K.H. Yoo and U. Gretzel. 2009. Comparison of Deceptive and Truthful Travel Reviews. *Information and Communication Technologies in Tourism 2009*, pages 37–47.

L. Zhou, J.K. Burgoon, D.P. Twitchell, T. Qin, and J.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.

L. Zhou, Y. Shi, and D. Zhang. 2008. A Statistical Language Modeling Approach to Online Deception Detection. *IEEE Transactions on Knowledge and Data Engineering*, 20(8):1077–1081.