Human Computation
AAAI Technical Report WS-12-08

Detecting Deceptive Opinion Spam Using Human Computation

Christopher G. Harris

Informatics Program, The University of Iowa, Iowa City, Iowa USA
christopher-harris@uiowa.edu

Abstract

Websites that encourage consumers to research, rate, and review products online have become an increasingly important factor in purchase decisions. This increased importance has been accompanied by a growth in deceptive opinion spam - fraudulent reviews written with the intent to sound authentic and mislead consumers. In this study, we pool deceptive reviews solicited through crowdsourcing with actual reviews obtained from product review websites. We then explore several human- and machine-based assessment methods to spot deceptive opinion spam in our pooled review set. We find that the combination of human-based assessment methods with easily-obtained statistical information generated from the review text outperforms detection methods using human assessors alone.

Introduction

Product reviews provide consumers, retailers, and manufacturers with information that impacts purchasing decisions. Consumers use these reviews not only to receive word-of-mouth (WOM) information on products, such as quality, suitability and utility, but also to provide input on their own product experience to other consumers. Retailers and manufacturers use these reviews to identify features that are important to consumers, which are then fed back into marketing and product development channels. A recent Cone Research study (Cone 2011) illustrates the power of online consumer reviews on purchases: 80% of consumers reverse product purchase decisions after reviewing negative consumer reviews, and 87% affirm a purchase decision based on positive consumer reviews.

Opinion spam, also called *review spam*, are reviews that range from simply annoying self-promotions or announcements that have no relationship with the reviewed product, to deliberately fraudulent product reviews provided with the intention of misleading consumers. This latter type is considered *deceptive opinion spam* and is the focus of our study. Deceptive opinion spam has two distinct variations: *hyper spam*, where unwarranted positive reviews are given to products in order to unfairly promote them, and *defaming spam*, which gives unjustified negative reviews to competing products in order to damage their reputations with consumers.

Although the true percentage of reviews containing deceptive spam is not known, some recent incidents have shed some light on how far spammers will go to rig product reviews. In early 2009, a manager at Belkin hired workers using Amazon Mechanical Turk (MTurk) to write positive reviews on a router suffering from poor product ratings (CNET, 2009). In late 2011, a leather case maker for the Kindle Fire reportedly offered a rebate on the entire purchase price in exchange for a five-star rating on Amazon (Streitfeld, 2012). This deception can also be costly to those involved with its creation. Orlando Figes, one of the U.K.'s leading historians on Russia, was ordered to pay libel damages in 2010 for posting defaming spam on rivals' books on Amazon (Topping, 2010). Legacy Learning Systems, a retailer of musical instruction DVDs, was recently fined $250,000 by the U.S. Federal Trade Commission after being charged for hiring affiliates to write positive reviews of their products on various websites (FTC, 2011). Due to the difficulty of opinion spam detection, these episodes likely represent only a tiny fraction of all such incidents.

In this paper, we provide the following contributions. First, we contrast two different human-based assessment methods using pooled sets of truthful product reviews and manually-created deceptive reviews. Second, we apply different types of assessment methods on both highly-rated (five-star) reviews and low-rated (one- and two-star) reviews, each of which display unique characteristics. Third, we compare human- and machine-based assessment methods for detecting deceptive opinion spam. In doing so, we develop a hybrid assessment design that significantly improves upon human-based spam-detection methods, without requiring the overhead of automated machine-based classification methods.

Related Work

A number of studies have focused on traditional spam detection in e-mail and on the web; however, only recently have any studies have examined opinion spam. Jindal and Liu (2008) performed some of the first studies of this nature. They focused on three types of disruptive opinion spam, including spam containing advertisements and other non-related text. While these types of spam may be distracting, they are easily detectable by human readers. In contrast, the focus of our study is detecting deceptive opinion spam, written with the specific intent of misleading customers and is therefore difficult for humans to detect (Bond and DePaulo, 2006). Opinion spam detection provides an unusual scenario in the assessment of human-created data, since machine-based methods have been shown to outperform human judges (Ott *et al*, 2011).

Jindal and Liu (2008) also examined duplicate (and near-duplicate) opinion spam, defined as the use of multiple reviews with similar text on the same website. Using the reviewer's profile, review text and product details, their study was able to detect duplicate (and near-duplicate) opinions. Although duplicate opinion spam can certainly influence product review ratings, a significant portion can be detected with freely-available plagiarism detection tools.

Yoo and Gretzel (2009) examined seven hypotheses on the linguistic dimensions of a set of 82 (40 truthful, 42 deceptive) highly-rated (five-star) TripAdvisor hotel reviews. Their analysis found deceptive reviews had greater lexical complexity, more frequent use of brand names, personal pronouns and words associated with positive sentiment than truthful reviews. In this study, we examine these same dimensions using our much larger product review dataset and compare their findings with ours.

Wu *et. al.* (2008) examined both hyper and defaming spam, but their focus was the distortion of user posting patterns on hotel reviews in TripAdvisor using a temporal analysis. Their methods rely on profiles of hotel reviewers and the hotel reviews received, whereas the focus in our study is expressly on the review text.

Hu *et. al.* (2012) combined sentiment mining techniques with readability assessments to detect deceptive opinion spam in Amazon book reviews. Their study found that 10.3% of the books examined were subject to some form of review manipulation. They also concluded that consumers can detect manipulation in ratings but not through review sentiment. We apply some of their methods in our study.

Ott *et. al.* (2011) also examined deceptive opinion spam in hotel reviews on TripAdvisor using a much larger dataset than Yoo and Gretzel. They developed several approaches involving text categorization, psycholinguistic deception detection, and identification of writing genre, developing an automatic classification technique that claims a nearly 90% accuracy on their class-balanced dataset. We follow many of their approaches in our study. One limitation of their study is the exclusive focus hyper spam – only examining five-star hotel reviews – while ignoring defaming spam.

A few studies have also examined the use of human computation methods in creating and detecting deceptive online spam. Harris (2011) and Wang *et al.* (2011) separately examined the difficulty in detecting fake online reviews created through crowdsourcing. Ghosh *et. al.* (2011) examined methods to moderate fake reviews using crowdsourcing techniques, indicating the crowdsourcing method scales well and are quickly adaptable to new threats.

Experimental Evaluation

The objective of this study is to examine how (hybrid) human computation methods stack up against both human-only assessment and automatic classification methods to detect deceptive opinion spam in both highly-rated and low-rated product reviews.

Data Preparation

Following the work of Yoo and Gretzel (2009 and Ott *et. al.* (2011), we begin with truthful product reviews and use the crowd to create fake product reviews. For truthful reviews, we obtained 2894 five-star (highly-rated) reviews, denoted $\boldsymbol{T^H}$ and 508 one- and two-star product (low-rated) reviews[1], denoted $\boldsymbol{T^L}$, on eight popular bodybuilding supplements from four sources: Amazon.com, bodybuilding.com GNC.com and supplementreviews.com[2].

We chose bodybuilding supplements because they show remarkably little product feature differentiation – most of the discussion contained in bodybuilding supplement reviews is focused on three product features: performance, price and taste. A limited set of product features allows us to examine specific linguistic qualities in detail. Information about the supplements examined is provided in Table 1.

[1] One- and two-star reviews were combined in $\boldsymbol{T^L}$ because the small number of one-star reviews meeting our criteria. A preliminary examination did not find any significant distinction between one-star and two-star reviews on any of the metrics we evaluated.

[2] bodybuilding.com and supplementreviews.com both use a ten-point rating scale. We use ratings of 9 and 10 for $\boldsymbol{T^H}$, and ratings 4 and below for $\boldsymbol{T^L}$.

Product Name	Mean rating (out of 5)	Nbr of T^H reviews available	Nbr of T^L reviews available
Optimum Nutrition 100% Whey Protein	4.256	988	102
BSN CellMass	4.229	464	74
BSN Syntha 6 Protein	4.151	428	72
Twinlab 100% Whey Protein Fuel	4.109	142	33
Dymatize Elite Whey Protein	4.096	241	53
Body Fortress Premium Whey Protein	4.088	226	61
Nature's Best Perfect Zero Carb isolate	4.042	263	64
BSN N.O. -Xplode	4.024	142	49
TOTAL	**4.124**	**2894**	**508**

Table 1: Products evaluated in this study, including the mean rating (on a five-point scale), and the number of available $\boldsymbol{T^H}$ and $\boldsymbol{T^L}$ reviews after unusable reviews are removed

		Review length, in words			
Set	**Count**	**Mean**	**Min**	**Max**	**Std. dev.**
$\boldsymbol{T^H}$	200	141.3	33	872	91.2
$\boldsymbol{T^L}$	200	102.7	29	438	66.3
$\boldsymbol{D^H}$	200	129.1	26	676	83.4
$\boldsymbol{D^L}$	200	93.4	22	309	61.8
ALL	**800**	**116.6**	**22**	**872**	**75.7**

Table 2: Metrics on each of our four review sets, including the minimum, maximum and average number of words and standard deviation.

Next, we removed reviews that contained fewer than 150 characters, were clearly off-topic, or that contained duplicates or near-duplicates of other reviews. From the remaining reviews, we randomly selected 25 reviews from $\boldsymbol{T^H}$ and 25 reviews from $\boldsymbol{T^L}$ for each of the eight products. We discarded the remaining reviews.

We then used MTurk to produce an equivalent number of both highly-rated and low-rated deceptive (fake) reviews for each product, denoted $\boldsymbol{D^H}$ and $\boldsymbol{D^L}$ respectively. Crowdworkers were provided with specific instructions on the minimum length (150 characters) and on the required polarity (positive or negative) of the review. We did not restrict the use of other resources in the preparation of their reviews; however, in order to avoid duplicates, we provided specific instructions to not plagiarize from existing reviews. We checked submissions using a web-based plagiarism detector[3]. Table 2 provides basic metrics for each of the four sets.

Human Assessor Measurements

Our next objective is to examine if human assessors can distinguish deceptive reviews and truthful reviews from a pooled set. We conduct this examination using workers hired from MTurk. In each case, decisions are determined by a simple majority voting method using three workers with instructions to make a determination only from an examination of the rating and the review text.

One assumption we make is that all bodybuilding supplement reviews collected from product websites are truthful, but this may not be the case. These actual website reviews, which we have marked as "truthful", may also be deceptive online spam. This could produce several false negatives undetectable by our study; however, this information is difficult to ascertain.

We examine two items in the study by Ott *et. al.* (2011). First, their study indicated the best inter-annotator agreement between any two of the three human judges was 0.12, indicating human detection methods were little better than chance. Second, they indicate the deception assessment ability of crowdworkers is inferior to the ability of student assessors. To examine this claim, we also asked three volunteer undergraduate student assessors, with no previous familiarity of bodybuilding supplements, to independently assess the same pooled reviews. We then applied a simple majority vote. Reviews were segregated by product and rating: $\boldsymbol{T^H}$ and $\boldsymbol{D^H}$ were pooled together to comprise a pooled group, $\boldsymbol{P^H}$, of highly-rated reviews,; likewise, $\boldsymbol{T^L}$ and $\boldsymbol{D^L}$ were pooled together to form a pooled set of low-rated reviews, $\boldsymbol{P^L}$, for each product). We explored the following two assessment scenarios on the two sets of pooled reviews:

> **Balanced.** Have assessors classify 5 truthful and 5 deceptive reviews each pooled set without the use of marked truthful and deceptive examples. Assessors are aware of the balanced ratio.
>
> **Random.** Provide 5 truthful and 5 deceptive reviews as labeled examples; randomly select *n* deceptive reviews, ($2 \leq n \leq 6$) and *10-n* truthful reviews for each product. Have assessors determine each from a pooled set. Assessors are not aware of the true ratio.

We believe the random scenario is more realistic, since the true mix of truthful and deceptive reviews is rarely known and there are often examples (e.g., reviews from a known source, such as an "editor's review"), which may be used as decision inputs. We repeated the balanced and random assessment scenarios for the pooled set of highly-rated reviews ($\boldsymbol{P^H}$) and the pooled set of low-rated reviews ($\boldsymbol{P^L}$) for each of the eight products. The results are provided in Table 3. This initial assessment demonstrates student assessors performed better than the crowd assessors, but this difference was not significant (one tailed

[3] http://plagiarisma.net/

High-rated reviews			Truthful (T^H)		Deceptive (D^H)	
Scenario	Assessor	Acc	P	R	P	R
Balanced	Students	.625	.596	.775	.679	.475
Balanced	Crowd	.638	.617	.725	.667	.550
Random	Students	.588	.696	.709	.333	.320
Random	Crowd	.575	.744	.582	.378	.560

Low-rated reviews			Truthful (T^L)		Deceptive (D^L)	
Scenario	Assessor	Acc	P	R	P	R
Balanced	Students	.438	.436	.425	.439	.450
Balanced	Crowd	.463	.500	.442	.486	.429
Random	Students	.450	.571	.549	.258	.276
Random	Crowd	.513	.623	.635	.296	.286

Table 3: Result from the human-based assessment task using students and the crowd, separated by review type and by scenario. We also report the accuracy (Acc), the precision (P) and recall (R) for each group.

sign test, p = 0.143). The inter-annotator agreement between human judges calculated using Fleiss' kappa (κ) was 0.24 in the balanced set, but drops to 0.16 in the random set. Likewise, inter-annotator agreement between the crowd-based judges drops from 0.19 for balanced set but drops to 0.14 for the random set[4]. There is no universal scale to interpreting these κ values, but in general larger values on the same judgment pairs imply greater decision confidence. In our case, it illustrates that detecting deceptive opinion spam is non-trivial; more importantly, it illustrates that the deceptive reviews were convincingly written by the crowd.

The assessors detected opinion spam in the P^H review set more easily than in the P^L review set (one-tailed sign test, p=0.03). This difference was significant for both balanced and random sets. We also note the low precision and recall scores for the deceptive opinion spam detection of the low-rated reviews indicate the difficulty of identifying fake reviews. This inability to recognize deceptive reviews may be a result of truth bias (Vrij, 1999, Elaad 2003) – a well-known condition in deception studies where an assessor has a "default" belief that a review must be true.

Writing Style Measurements

We examined three linguistic qualities for each review – polarity, sentiment, and readability. These components are relatively quick and inexpensive to calculate and serve as inputs into our automatic classification section (discussed in the next section).

To analyze the sentiment, we use the sentiment API provided by text-processing.com[5]. Although trained on movie reviews, the sentiment analyzer performed well on earlier empirical tests on our data. This tool uses a hierarchical classification approach; neutrality is initially decided, then sentiment polarity is determined on the polarized text. Polarity/neutrality and positive/negative sentiment are scored on separate (0,1) scales. We also ascertain the complexity of each review using the Automated Readability Index (ARI)[6]. The ARI decomposes text into its structural elements to provide the minimum reading level needed to understand a snippet of text, based on United States grade levels. The formula used to calculate ARI is given as:

$$\mathrm{ARI} = 4.71(C/W) + 0.5\,(W/S) - 21.43$$

where C, W and S are the total number of characters, words, and sentences contained in each review, respectively. Unlike other readability metrics, the ARI metric takes into account the number of characters, instead of syllables, in each word. In their study, Hu *et. al.* (2012) determined deceptive and truthful reviews often display different ARI distributions. We wish to see if this also applies with the 800 reviews used in our study. Summary information is provided in Table 4.

Set	ARI	Polarity	Sentiment
T^H	7.96	0.79	0.86
T^L	6.86	0.84	0.88
D^H	7.14	0.88	0.90
D^L	5.79	0.93	0.91
ALL	**6.94**	**0.86**	**0.88**

Table 4: Readability, polarity and sentiment scores for each review set. ARI reflects the minimum U.S. grade level for comprehension, polarity indicates deviation from neutrality, and sentiment indicates positive or negative word alignment.

In Table 4, we examine the distribution of sentiment and ARI scores. As expected, the sentiment scores are clearly polarized for P^H and P^L sets, but the overall difference in sentiment between truthful and deceptive reviews is not significant for polarity (two-tailed t-test, polarity: p= 0.085, 0.101) or sentiment (two-tailed t-test p=0.053, 0.074) for highly and low rated reviews, respectively. ARI scores appear to give a better indicator. Readability scores show a significant difference between the P^H and P^L sets, with P^H sets using language requiring higher readability (two-tailed t-test, p<0.001, p<0.001). Graphs illustrating the score distribution make the difference more easily

[4] Inter-group ratings, using majority vote of the student group and the crowd group, are as follows: highly-rated balanced 0.70, highly-rated random: 0.64, low-rated balanced: 0.72, low-rated random: 0.69, indicating a strong agreement between sets on both sets using the two assessment methods.

[5] http://text-processing.com/api/sentiment/

[6] An online calculator can be found at the following URL: http://www.online-utility.org/english/readability_test_and_improve.jsp

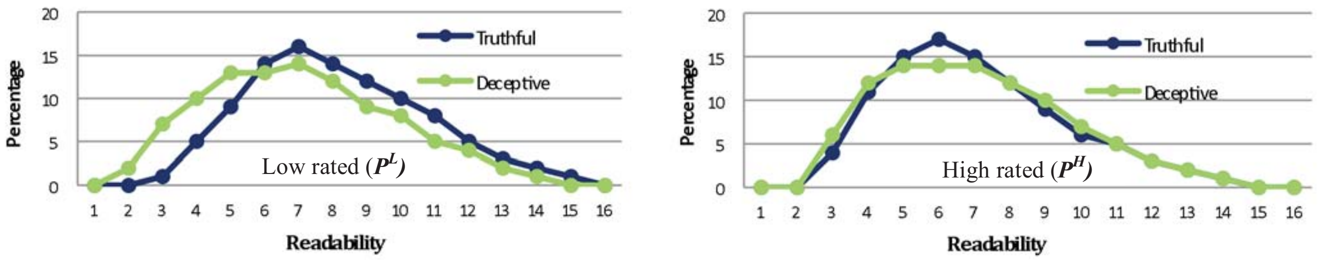

Figure 1: Readability comparison between truthful and deceptive language in low-rated (left) and high-rated (right) reviews using ARI.

detectable – for both sets, we observe a lower, flatter distribution for the deceptive reviews (see Figure 1).

We now examine the seven linguistic dimensions mentioned by Yoo and Gretzel (2009) in their study on TripAdvisor reviews. They examined the differences in the length of the review based on word *quantity* (refer to Table 2 for our data) and found no difference; likewise, we found no difference (two-tailed t-test, p=0.072, 0.081 for highly- and low-rated sets, respectively).

We now examine the seven linguistic dimensions mentioned by Yoo and Gretzel (2009) in their study on truthful and deceptive TripAdvisor reviews. They examined the differences in the length of the review based on word *quantity* (refer to Table 2 for our data) and found no difference; likewise we found no difference (two-tailed t-test, p=0.072, 0.081 for highly- and low-rated sets, respectively). They also examined the *complexity* in deceptive reviews and found a greater complexity with deceptive reviews. We found the opposite to be true based on ARI (Table 4). Next, they examined word *diversity* (the ratio of unique words to total words); no significant difference was found by them or by us (two-tailed t-test, p=0.09, 0.11 for highly- and low- rated sets, respectively) and *immediacy* (the ratio of number of first person pronouns to total words). They discovered deceptive reviews are significantly more likely to include self-references; our results indicated a significant difference in the same direction (deceptive reviews contain more self-references in both highly- and low-rated reviews (two-tailed t-test, p=0.013, 0.022 for highly- and low-rated sets, respectively).

Dimension	Description	Yoo & Gretzel Finding	Finding on our data

Quantity	Total number of words	No significant difference	No diff
			p = 0.072
			p = 0.081

Complexity	Writing level	Deceptive higher	Truthful higher p < 0.001 p < 0.001

Diversity	Number of unique words	No significant difference	No diff
			p = 0.093
			p = 0.111

Immediacy	Ratio of first person pronouns to total words	Deceptive higher	Deceptive higher
			p = 0.013
			p = 0.022

Branding	Percent of reviews mentioning brand	Deceptive higher	Deceptive higher p = 0.024 p < 0.001

Positive Sentiment	Positive sentiment score	Deceptive higher	No difference on sentiment
			p = 0.053
			p = 0.074

Negative Sentiment	Negative sentiment score	Truthful higher	

Table 5: An examination of the findings from the Yoo and Gretzel review study on TripAdvisor ratings and the comparable findings on our review set. The first p-value listed is for highly-rated pooled review set (***P^H*)**, second is for the low-rated pooled review set (***P^L*).**

Another dimension they examined was *branding*, which examined the number of times the brand was mentioned in a review. They found deceptive reviews repeated the brand name significantly more often; indeed, our examination also showed a more frequent mention of the brand name in deceptive reviews, particularly for the high-rated review set (two-tailed t-test, p=0.024, 0.001 for highly- and low rated sets, respectively). Finally, they found their deceptive reviews displayed *positive sentiment* more often while truthful reviews displayed *negative sentiment* more often; our results (refer to Table 4 for our data) found no significant differences in either set of reviews. Table 5 summarizes Yoo and Gretzel's findings on their reviews, as well as the equivalent findings on our reviews. We do note that, on some evaluations such as sentiment and complexity, the measurements are made using different metrics; in addition, the review domain is different (hotels vs. supplement reviews) and the number of reviews evaluated was different (82 vs. 800).

Classifier Measurements

We obtain two balanced review sets: ***P^H***, comprised of 320 ***T^H*** and 320 ***D^H*** reviews[7], and another of equal size

[7] We excluded the 5 deceptive and 5 truthful reviews per product used as examples in the earlier human assessment step in our classification model

comprised of the $\boldsymbol{P^L}$ reviews. Using the QuickLM language model toolkit[8], we create two separate unigram (lower case and unstemmed) language models, one for each set of pooled reviews, $\boldsymbol{P^H}$ and $\boldsymbol{P^L}$. We use each language model, along with sentiment scores and ARI determined earlier, as feature set inputs to SVM$^{\text{light}}$ (Joachims 1999). We use interpolated Kneser-Ney smoothing (Chen, 1999) and normalize document vectors to unit-length. For simplicity, we restrict our evaluation to linear-kernel SVMs, which classify a document $\vec{x}$ using a learned weight vector $\vec{w}$ and bias term b using the following equation:

$$\mathit{class} = \text{sign}\langle \vec{x}, \vec{w} \rangle + b$$

where $\langle \vec{x}, \vec{w} \rangle$ is the vector dot product. Following the method suggested by Quadrianto (2003) and used by Ott *et. al.* (2011), we do a five-fold cross validation on each model and set up our model so each fold comprises either all highly- or all low-rated reviews for two supplement products. This ensures learned models are always evaluated on unseen products in each fold. The results are provided in Table 6.

		Truthful		Deceptive	
Review set	**Acc**	**P**	**R**	**P**	**R**
High (balanced)	.763	.775	.756	.769	.750
Low (balanced)	.738	.686	.875	.828	.600

Table 6: Result from the automatic classification methods for the two review sets: the high-rated reviews ($\boldsymbol{P^H}$) and the low-rated reviews ($\boldsymbol{P^L}$).

In Table 6, we observe that our automated classifier significantly outperformed our human and crowd assessors; however, we were unable to achieve the impressive unigram accuracy numbers Ott *et. al.* (2011) did with on TripAdvisor hotel reviews. Detecting deceptive opinion spam within the $\boldsymbol{P^H}$ set was easier using this model than within the $\boldsymbol{P^L}$ set, but this difference was not significant (one tailed sign test, p=0.081). Both were significant improvements over the human assessor results (one tailed sign test, $p < 0.001$, $p < 0.001$).

Hybrid Measurements

Next, we wish to examine if the linguistic properties, sentiment scores and ARI scores calculated for each review could empower human and crowd assessors to match the scores obtained by our automatic classifier. Using the same three student assessors, a new set of MTurk assessors and an equivalent number of reviews not used in the first human assessment, we repeated the same two assessment scenarios – however, this time we provide some additional information: the sentiment, ARI, and Yoo and Gretzel dimension scores for each review as well as the product means for each review set (e.g., deceptive ARI scores are significantly higher than truthful ARI scores for low-rated reviews). The results for human assessors using this additional data are provided in Table 7.

High-rated reviews with additional data			**Truthful (T^H)**		**Deceptive (D^H)**	
Scenario	**Assessor**	**Acc**	**P**	**R**	**P**	**R**
Balanced	Students	.688	.674	.725	.703	.650
Balanced	Crowd	.663	.638	.750	.697	.575
Random	Students	.650	.829	.618	.462	.720
Random	Crowd	.688	.788	.745	.500	.560

Low-rated reviews with additional data			**Truthful (T^L)**		**Deceptive (D^L)**	
Scenario	**Assessor**	**Acc**	**P**	**R**	**P**	**R**
Balanced	Students	.575	.565	.650	.588	.500
Balanced	Crowd	.563	.568	.525	.558	.600
Random	Students	.575	.659	.574	.487	.576
Random	Crowd	.600	.696	.667	.500	.533

Table 7: Result from the human-based assessment task using students and the crowd with additional measurement data, separated by review type and by scenario.

From Table 7, we see a significant improvement over the results obtained from Table 3, particularly in the low-rated reviews. Although the accuracy of our human judgments with additional statistical information on the review is still lower than that obtained using the automated classifier (one tailed sign test, p=0.042, p=0.035), a combination of human- and machine-based assessment tools outperforms human assessment alone (one tailed sign test, $p<0.001$, $p<0.001$). Therefore, providing human and crowd assessors with meaningful metrics is likely to improve the quality on other assessment tasks as well, with relatively little cost.

Table 8 illustrates five of the highest-weighted keywords in our language models for both truthful and deceptive reviews. We observe that most of the truthful reviews discuss the product (e.g., "casein", "release", "mix") and how it affects the workout performance (e.g., "pumps", "energy", "release"), whereas keywords associated with the deceptive reviews are generic and could be associated with reviews on any topic (e.g., "scam" "waste", "stuff").

Highly-Rated		**Low-Rated**	
(T^H)	***(D^H)***	***(T^L)***	***(D^L)***
good energy release casein pumps	drink delicious quality stuff gym	mix scoops crash poop pricey	waste back scam money fart

Table 8: Top 5 unigrams detected by our language model, broken into groups by rating type (high or low) and by source (truthful or deceptive).

[8] http://www.speech.cs.cmu.edu/tools/lm.html

Since the linguistic feature statistics provided to our human assessors were relatively quick and inexpensive to calculate (as compared with preparing, training, and evaluating using an automatic classifier), we believe providing human assessors with additional statistical information can make a significant improvement to human assessment at very little cost. However, this low cost could benefit spammers as well; the statistical information should not be recklessly disclosed, as spammers can also adapt by utilizing the same information when preparing fake product reviews. With the large number of linguistic features available to monitor, coupled with the different levels of granularity at which they may be evaluated, reproduction could be hindered if a weighted formula used for assessment was changed frequently.

Although this study shows automatic classification methods still outperform human-based assessment methods, there are ways to narrow this gap at little cost. Some situations may make automatic methods impractical to implement; for example, when there are extensive comparisons between two products in a single review, it can confuse the automatic classifier. In other cases, the classifier may not be able to come up with a set of labeled examples (e.g., a small number of samples). Therefore, providing human assessors with low-cost metadata on the linguistic properties of review datasets is one way to improve the ability to detect deceptive opinion spam in product reviews.

Conclusion

In this study, we created a dataset comprised of 400 deceptive reviews and 400 actual reviews. We used this data to examine a variety of human-based, machine-based, and hybrid assessment methods to detect deceptive opinion spam in product reviews. We examined both hyper-spam (on highly-rated reviews) and defaming spam (on low-rated reviews), each of which has a unique set of assessment challenges that have not been previously investigated in the literature.

Future efforts will be targeted on analyzing the data and exploring new uses for detecting deceptive opinion spam. The data collected for this study will be used for additional studies involving human computation in recommender systems. We plan to examine if using crowdworkers who are familiar with bodybuilding supplements can outperform the automatic classifier, given this additional statistical information about the product reviews.

References

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

Chen, S. F. and Goodman, J. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13:4 1999, p 359-393.

CNET. "Fake reviews prompt belkin apology". Available at: http://news.cnet.com/8301-1001_3-10145399-92.html. 2009

Cone Research. "Game changer: cone survey finds 4-out-of-5 consumers reverse purchase decisions based on negative online reviews". Available at: http://www.coneinc.com/negative-reviews-online-reverse-purchase-decisions 2011.

Elaad, E. Effects of feedback on the overestimated capacity to detect lies and the underestimated ability to tell lies. *Applied Cognitive Psychology*, 17:3 2003, p 349-363.

FTC. "Firm to pay ftc $250,000 to settle charges that it used misleading online 'consumer' and 'independent' reviews". Available at: http://www.ftc.gov/opa/2011/03/legacy.shtm. 2011.

Ghosh, A., Kale, S. and McAfee, P. Who Moderates the Moderators? Crowdsourcing Abuse Detection in User-Generated Content. In *Proc of the 12th ACM conference on Electronic commerce* (EC '11). ACM, New York, NY, 167-176.

Harris, C. G. Dirty Deeds Done Dirt Cheap: A Darker Side to Crowdsourcing. In *Proc of IEEE SocialCom'11*, Boston, MA. 2011. 1314-1317

Hu, N., Bose, I., Koh, N. S. and Liu, L. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems,* 2011.

Jindal, N. and Liu, B. *Opinion spam and analysis*. In Proc. WSDM '08. ACM, New York, NY, 2008, 219-230.

Joachims, T. Making large scale SVM learning practical.1999.

Ott, M., Choi, Y., Cardie, C. and Hancock, J. T. Finding deceptive opinion spam by any stretch of the imagination. In *Proc ACL-HLT '11, Vol. 1*. ACL, Stroudsburg, PA, 309-319.

Quadrianto, N., Smola, A. J., Caetano, T. S. and Le, Q. V. Estimating labels from label proportions. *J. Mach. Learn. Res.* 10 (December 2009), 2349-2374.

Streitfeld, D. "For $2 a star, an online retailer gets 5-star product reviews". New York Times. January 26, 2012. Available at: http://www.nytimes.com/2012/01/27/technology/for-2-a-star-a-retailer-gets-5-star-reviews.html. 2012.

Topping, A. Historian orlando figes agrees to pay damages for fake reviews. The Guardian. July 16, 2010. Available at: http://www.guardian.co.uk/books/2010/jul/16/orlando-figes-fake-amazon-reviews. 2010.

Vrij, A. and Baxter, M. Accuracy and confidence in detecting truths andlies in elaborations and denials: Truth bias, lie bias and individual differences. *Expert evidence*, 7:1 1999, 25-36.

Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H. and Zhao, B. Y. Serf and Turf: Crowdturfing for Fun and Profit. In *Proc. WWW '12*. ACM, New York, NY, 679-688.

Wu, G., Greene, D., Smyth, B. and Cunningham, P. Distortion as a validation criterion in the identification of suspicious reviews. In *Proc SOMA '10*. ACM, New York, NY, 10-13.

Yoo, K. H. and Gretzel, U. Comparison of Deceptive and Truthful Travel Reviews. *Information and communication technologies in tourism, 2009,* 37-47.