Negative Deceptive Opinion Spam

Myle Ott **Claire Cardie**
Department of Computer Science
Cornell University
Ithaca, NY 14853

Jeffrey T. Hancock
Department of Communication
Cornell University
Ithaca, NY 14853

{myleott,cardie}@cs.cornell.edu jeff.hancock@cornell.edu

Abstract

The rising influence of user-generated online reviews (Cone, 2011) has led to growing incentive for businesses to solicit and manufacture DECEPTIVE OPINION SPAM—fictitious reviews that have been deliberately written to sound authentic and deceive the reader. Recently, Ott et al. (2011) have introduced an opinion spam dataset containing gold standard deceptive *positive* hotel reviews. However, the complementary problem of *negative* deceptive opinion spam, intended to slander competitive offerings, remains largely unstudied. Following an approach similar to Ott et al. (2011), in this work we create and study the first dataset of deceptive opinion spam with *negative* sentiment reviews. Based on this dataset, we find that standard n-gram text categorization techniques can detect negative deceptive opinion spam with performance far surpassing that of human judges. Finally, in conjunction with the aforementioned *positive* review dataset, we consider the possible interactions between sentiment and deception, and present initial results that encourage further exploration of this relationship.

1 Introduction

Consumer's purchase decisions are increasingly influenced by user-generated online reviews of products and services (Cone, 2011). Accordingly, there is a growing incentive for businesses to solicit and manufacture DECEPTIVE OPINION SPAM—fictitious reviews that have been deliberately written to sound authentic and deceive the reader (Ott et al., 2011). For example, Ott et al. (2012) has estimated that between 1% and 6% of *positive* hotel reviews appear to be deceptive, suggesting that some hotels may be posting fake positive reviews in order to hype their own offerings.

In this work we distinguish between two kinds of deceptive opinion spam, depending on the sentiment expressed in the review. In particular, reviews intended to promote or hype an offering, and which therefore express a positive sentiment towards the offering, are called *positive* deceptive opinion spam. In contrast, reviews intended to disparage or slander competitive offerings, and which therefore express a negative sentiment towards the offering, are called *negative* deceptive opinion spam. While previous related work (Ott et al., 2011; Ott et al., 2012) has explored characteristics of *positive* deceptive opinion spam, the complementary problem of *negative* deceptive opinion spam remains largely unstudied.

Following the framework of Ott et al. (2011), we use Amazon's Mechanical Turk service to produce the first publicly available[1] dataset of *negative* deceptive opinion spam, containing 400 gold standard deceptive negative reviews of 20 popular Chicago hotels. To validate the credibility of our deceptive reviews, we show that human deception detection performance on the negative reviews is low, in agreement with decades of traditional deception detection research (Bond and DePaulo, 2006). We then show that standard n-gram text categorization techniques can be used to detect negative deceptive opinion spam with approximately 86% accuracy — far

[1]Dataset available at: http://www.cs.cornell.edu/~myleott/op_spam.

Proceedings of NAACL-HLT 2013, pages 497–501,
Atlanta, Georgia, 9–14 June 2013.

surpassing that of the human judges.

In conjunction with Ott et al. (2011)'s *positive* deceptive opinion spam dataset, we then explore the interaction between sentiment and deception with respect to three types of language features: (1) changes in first-person singular use, often attributed to psychological distancing (Newman et al., 2003), (2) decreased spatial awareness and more narrative form, consistent with theories of reality monitoring (Johnson and Raye, 1981) and imaginative writing (Biber et al., 1999; Rayson et al., 2001), and (3) increased negative emotion terms, often attributed to leakage cues (Ekman and Friesen, 1969), but perhaps better explained in our case as an exaggeration of the underlying review sentiment.

2 Dataset

One of the biggest challenges facing studies of deception is obtaining labeled data. Recently, Ott et al. (2011) have proposed an approach for generating *positive* deceptive opinion spam using Amazon's popular Mechanical Turk crowdsourcing service. In this section we discuss our efforts to extend Ott et al. (2011)'s dataset to additionally include *negative* deceptive opinion spam.

2.1 Deceptive Reviews from Mechanical Turk

Deceptive negative reviews are gathered from Mechanical Turk using the same procedure as Ott et al. (2011). In particular, we create and divide 400 HITs evenly across the 20 most popular hotels in Chicago, such that we obtain 20 reviews for each hotel. We allow workers to complete only a single HIT each, so that each review is written by a unique worker.[2] We further require workers to be located in the United States and to have an average past approval rating of at least 90%. We allow a maximum of 30 minutes to complete the HIT, and reward accepted submissions with one US dollar ($1).

Each HIT instructs a worker to imagine that they work for the marketing department of a hotel, and that their manager has asked them to write a fake negative review of a competitor's hotel to be posted online. Accompanying each HIT is the name and URL of the hotel for which the fake negative review is to be written, and instructions that: (1) workers should not complete more than one similar HIT, (2) submissions must be of sufficient quality, i.e., written for the correct hotel, legible, reasonable in length,[3] and not plagiarized,[4] and, (3) the HIT is for academic research purposes.

Submissions are manually inspected to ensure that they are written for the correct hotel and to ensure that they convey a generally negative sentiment.[5] The average accepted review length was 178 words, higher than for the positive reviews gathered by Ott et al. (2011), who report an average review length of 116 words.

2.2 Truthful Reviews from the Web

Negative (1- or 2-star) truthful reviews are mined from six popular online review communities: Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor, and Yelp. While reviews mined from these communities cannot be considered *gold standard* truthful, recent work (Mayzlin et al., 2012; Ott et al., 2012) suggests that deception rates among travel review portals is reasonably small.

Following Ott et al. (2011), we sample a subset of the available truthful reviews so that we retain an equal number of truthful and deceptive reviews (20 each) for each hotel. However, because the truthful reviews are on average longer than our deceptive reviews, we sample the truthful reviews according to a log-normal distribution fit to the lengths of our deceptive reviews, similarly to Ott et al. (2011).[6]

3 Deception Detection Performance

In this section we report the deception detection performance of three human judges (Section 3.1) and supervised n-gram Support Vector Machine (SVM) classifiers (Section 3.2).

[2]While Mechanical Turk does not provide a convenient mechanism for ensuring the uniqueness of workers, this constraint can be enforced with Javascript. The script is available at: `http://uniqueturker.myleott.com`.

[3]We define "reasonable length" to be ≥ 150 characters.

[4]We use `http://plagiarisma.net` to determine whether or not a review is plagiarized.

[5]We discarded and replaced approximately 2% of the submissions, where it was clear that the worker had misread the instructions and instead written a deceptive *positive* review.

[6]We use the R package GAMLSS (Rigby and Stasinopoulos, 2005) to fit a log-normal distribution (left truncated at 150 characters) to the lengths of the deceptive reviews.

			TRUTHFUL			DECEPTIVE		
		Accuracy	**P**	**R**	**F**	**P**	**R**	**F**
HUMAN	JUDGE 1	65.0%	65.0	65.0	65.0	65.0	65.0	65.0
	JUDGE 2	61.9%	63.0	57.5	60.1	60.9	66.3	63.5
	JUDGE 3	57.5%	57.3	58.8	58.0	57.7	56.3	57.0
META	MAJORITY	**69.4%**	70.1	**67.5**	**68.8**	**68.7**	71.3	**69.9**
	SKEPTIC	58.1%	**78.3**	22.5	35.0	54.7	**93.8**	69.1

Table 1: Deception detection performance, incl. (P)recision, (R)ecall, and (F)1-score, for three human judges and two meta-judges on a set of 160 *negative* reviews. The largest value in each column is indicated with boldface.

3.1 Human Performance

Recent large-scale meta-analyses have shown human deception detection performance is low, with accuracies often not much better than chance (Bond and DePaulo, 2006). Indeed, Ott et al. (2011) found that two out of three human judges were unable to perform statistically significantly better than chance (at the $p < 0.05$ level) at detecting *positive* deceptive opinion spam. Nevertheless, it is important to subject our reviews to human judgments to validate their convincingness. In particular, if human detection performance is found to be very high, then it would cast doubt on the usefulness of the Mechanical Turk approach for soliciting gold standard deceptive opinion spam.

Following Ott et al. (2011), we asked three volunteer undergraduate university students to read and make assessments on a subset of the negative review dataset described in Section 2. Specifically, we randomized all 40 deceptive and truthful reviews from each of four hotels (160 reviews total). We then asked the volunteers to read each review and mark whether they believed it to be truthful or deceptive.

Performance for the three human judges appears in Table 1. We additionally show the deception detection performance of two meta-judges that aggregate the assessments of the individual human judges: (1) the MAJORITY meta-judge predicts *deceptive* when at least two out of three human judges predict *deceptive* (and *truthful* otherwise), and (2) the SKEPTIC meta-judge predicts *deceptive* when at least one out of three human judges predicts *deceptive* (and *truthful* otherwise).

A two-tailed binomial test suggests that JUDGE 1 and JUDGE 2 both perform better than chance ($p = 0.0002, 0.003$, respectively), while JUDGE 3 fails to reject the null hypothesis of performing at-chance ($p = 0.07$). However, while the best human judge is accurate 65% of the time, inter-annotator agreement computed using Fleiss' kappa is only *slight* at 0.07 (Landis and Koch, 1977). Furthermore, based on Cohen's kappa, the highest pairwise inter-annotator agreement is only 0.26, between JUDGE 1 and JUDGE 2. These low agreements suggest that while the judges may perform statistically better than chance, they are identifying different reviews as deceptive, i.e., few reviews are consistently identified as deceptive.

3.2 Automated Classifier Performance

Standard n-gram–based text categorization techniques have been shown to be effective at detecting deception in text (Jindal and Liu, 2008; Mihalcea and Strapparava, 2009; Ott et al., 2011; Feng et al., 2012). Following Ott et al. (2011), we evaluate the performance of linear Support Vector Machine (SVM) classifiers trained with unigram and bigram term-frequency features on our novel *negative* deceptive opinion spam dataset. We employ the same 5-fold stratified cross-validation (CV) procedure as Ott et al. (2011), whereby for each cross-validation iteration we train our model on all reviews for 16 hotels, and test our model on all reviews for the remaining 4 hotels. The SVM cost parameter, C, is tuned by nested cross-validation on the training data.

Results appear in Table 2. Each row lists the sentiment of the train and test reviews, where "Cross Val." corresponds to the cross-validation procedure described above, and "Held Out" corresponds to classifiers trained on reviews of one sentiment and tested on the other. The results suggest that n-gram–based SVM classifiers can detect *negative* deceptive opinion spam in a balanced dataset with performance far surpassing that of untrained human judges (see Section 3.1). Furthermore, our results show that

			TRUTHFUL			DECEPTIVE		
Train Sentiment	**Test Sentiment**	**Accuracy**	**P**	**R**	**F**	**P**	**R**	**F**
POSITIVE (800 reviews)	POSITIVE (800 reviews, Cross Val.)	89.3%	89.6	88.8	89.2	88.9	89.8	89.3
	NEGATIVE (800 reviews, Held Out)	75.1%	69.0	91.3	78.6	87.1	59.0	70.3
NEGATIVE (800 reviews)	POSITIVE (800 reviews, Held Out)	81.4%	76.3	91.0	83.0	88.9	71.8	79.4
	NEGATIVE (800 reviews, Cross Val.)	86.0%	86.4	85.5	85.9	85.6	86.5	86.1
COMBINED (1600 reviews)	POSITIVE (800 reviews, Cross Val.)	88.4%	87.7	89.3	88.5	89.1	87.5	88.3
	NEGATIVE (800 reviews, Cross Val.)	86.0%	85.3	87.0	86.1	86.7	85.0	85.9

Table 2: Automated classifier performance for different train and test sets, incl. (P)recision, (R)ecall and (F)1-score.

classifiers trained and tested on reviews of different sentiments perform worse, despite having more training data,[7] than classifiers trained and tested on reviews of the same sentiment. This suggests that cues to deception differ depending on the sentiment of the text (see Section 4).

Interestingly, we find that training on the combined sentiment dataset results in performance that is comparable to that of the "same sentiment" classifiers (88.4% vs. 89.3% accuracy for positive reviews and 86.0% vs. 86.0% accuracy for negative reviews). This is explainable in part by the increased training set size (1,280 vs. 640 reviews per 4 training folds).

4 Interaction of Sentiment and Deception

An important question is how language features operate in our fake negative reviews compared with the fake positive reviews of Ott et al. (2011). For example, fake positive reviews included less spatial language (e.g., floor, small, location, etc.) because individuals who had not actually experienced the hotel simply had less spatial detail available for their review (Johnson and Raye, 1981). This was also the case for our negative reviews, with less spatial language observed for fake negative reviews relative to truthful. Likewise, our fake negative reviews had more verbs relative to nouns than truthful, suggesting a more narrative style that is indicative of imaginative writing (Biber et al., 1999; Rayson et al., 2001), a pattern also observed by Ott et al. (2011).

There were, however, several important differences in the deceptive language of fake negative relative to fake positive reviews. First, as might be expected, negative emotion terms were more frequent, according to LIWC (Pennebaker et al., 2007), in our fake negative reviews than in the fake positive reviews. But, importantly, the fake negative reviewers over-produced negative emotion terms (e.g., terrible, disappointed) relative to the truthful reviews in the same way that fake positive reviewers over-produced positive emotion terms (e.g., elegant, luxurious). Combined, these data suggest that the more frequent negative emotion terms in the present dataset are not the result of "leakage cues" that reveal the emotional distress of lying (Ekman and Friesen, 1969). Instead, the differences suggest that fake hotel reviewers exaggerate the sentiment they are trying to convey relative to similarly-valenced truthful reviews.

Second, the effect of deception on the pattern of pronoun frequency was not the same across positive and negative reviews. In particular, while first person singular pronouns were produced more frequently in fake reviews than truthful, consistent with the case for positive reviews, the increase was diminished in the negative reviews examined here. In the positive reviews reported by Ott et al. (2011), the rate of first person singular in fake reviews (M=4.36%, SD=2.96%) was twice the rate observed in truthful reviews (M=2.18%, SD=2.04%). In contrast, the rate of first person singular in the deceptive negative reviews (M=4.47%, SD=2.83%) was only 57% greater than for truthful reviews (M=2.85%, SD=2.23%). These results suggest that the emphasis on the self, perhaps as a strategy of convincing the reader that the author had actually been to the hotel, is not as evident in the fake negative reviews, perhaps because the negative tone of the reviews caused the reviewers to psychologically distance themselves from their negative statements, a phenomenon observed in several other deception studies, e.g., Hancock et al. (2008).

[7] "Cross Val." classifiers are effectively trained on 80% of the data and tested on the remaining 20%, whereas "Held Out" classifiers are trained and tested on 100% of each data.

5 Conclusion

We have created the first publicly-available corpus of gold standard *negative* deceptive opinion spam, containing 400 reviews of 20 Chicago hotels, which we have used to compare the deception detection capabilities of untrained human judges and standard n-gram–based Support Vector Machine classifiers. Our results demonstrate that while human deception detection performance is greater for *negative* rather than *positive* deceptive opinion spam, the best detection performance is still achieved through automated classifiers, with approximately 86% accuracy.

We have additionally explored, albeit briefly, the relationship between sentiment and deception by utilizing Ott et al. (2011)'s positive deceptive opinion spam dataset in conjunction with our own. In particular, we have identified several features of language that seem to remain consistent across sentiment, such as decreased awareness of spatial details and exaggerated language. We have also identified other features that vary with the sentiment, such as first person singular use, although further work is required to determine if these differences may be exploited to improve deception detection performance. Indeed, future work may wish to jointly model sentiment and deception in order to better determine the effect each has on language use.

Acknowledgments

This work was supported in part by NSF Grant BCS-0904822, a DARPA Deft grant, the Jack Kent Cooke Foundation, and by a gift from Google. We also thank the three Cornell undergraduate volunteer judges, as well as the NAACL reviewers for their insightful comments, suggestions and advice on various aspects of this work.

References

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.

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

P. Ekman and W.V. Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88.

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.

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.

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.

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

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

Dina Mayzlin, Yaniv Dover, and Judith A Chevalier. 2012. Promotional reviews: An empirical investigation of online review manipulation. Technical report, National Bureau of Economic Research.

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.

M. Ott, Y. Choi, C. Cardie, and J.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-Volume 1*, pages 309–319. Association for Computational Linguistics.

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.

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 (www.liwc.net)*.

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,(with discussion). *Applied Statistics*, 54:507–554.