Automatic detection of deceptive opinions using automatically identified specific details

Nikolai Vogler

April 10, 2017

1 Introduction

1.1 Deceptive Opinions

Distinguishing deceptive opinions — that is, fabricated views disguised to be genuine — from honest opinions is a hard problem. Deceptive opinions can include things like the false expression of a controversial opinion, a misleading review of an item or service bought online, or deceitful interviews. Unlike many tasks involving language, detecting deceptive opinions through text alone turns out to be quite difficult for humans. Ott et al. (2011) demonstrated this by asking a group of students to judge whether 40 online reviews were truthful or deceptive. These reviews were drawn from their Deceptive Opinion Spam Corpus, introduced in the same paper, and so included an equal number of truthful and deceptive reviews. The students performed at roughly 60% accuracy — only slightly better than the baseline, chance performance of selecting one of two choices (truthful or deceptive), which is 50%. Notably, the students were psychologically biased towards judging more opinions as truthful rather than deceptive. This poor performance suggests that detection of deceptive opinions is a complex area that can greatly benefit from unbiased computational analysis.

Much of the research performed on deceptive opinions has used online reviews from Ott et al. (2011)'s corpus as a benchmark because these data are a rich source of opinion spam, a type of deceptive opinion. As e-commerce burgeons, online reviews are becoming increasingly important to company reputations and consumer product assessment. Due to its influential impact on potential customers, deceptive opinion spam is being produced to deceive potential consumers. Whether the opinion spam is sponsored by a company wanting to promote its services (positive opinion spam) or a business maligning its rival with false claims (negative opinion spam), accurate

detection of opinion spam would prove extremely useful for both companies and consumers. Moreover, an automated process that could classify truthful and deceptive opinions more effectively than humans could effortlessly be run on large datasets. Not only could it be used to detect deceptive product reviews but also deceptive expressions of opinions more generally (e.g., controversial opinions, personal details during job interviews).

1.2 Purpose

Ott et al. (2011) noted informally that many truthful hotel reviews seemed to include specific details, such as specific spatial configurations in a hotel room. While this detail is clearly not generally appropriate for all opinions (unless they're about rooms), the idea that specific details are a hallmark of truth is something worth investigating.

In fact, we can see why specific details are important by examining Ott et al. (2013)'s dataset and identifying how specific detail features could potentially aid in classification performance of a truthful hotel review.

> The James Hotel met and exceeded our expectations. A haven of **cool, uncluttered comfort in a hot, crowded but congenial city**. **Located within steps of some of the greatest art, architecture, culinary and cultural opportunities anywhere in North America**, the staff of the James provided us with a home away from home for the week we visited. **The check-in was quick and flawless** and **the king suite a calm, restorative oasis**. It is difficult to over-praise the friendly, attentive staff; this is a hotel that truly has its act together! **The day before our visit ended, we experienced problems with the room's air conditioner**. When the technician was unable to quickly solve the problem, we were immediately upgraded to a **one-bedroom apartment with stunning views of the Chicago skyline**. We eagerly await the opening of new James Hotels in other cities.

In the above review of the James Hotel, many specific details are bolded in order to demonstrate their abundance. We can easily see that it is difficult to capture such details with only shallow, word-level features. Thus, we propose using more sophisticated phrase-level features to capture their presence.

2 Background

2.1 Datasets

We investigate deceptive opinions occurring in three separate datasets, which represent a broad sampling of deceptive opinions across the domains of online hotel reviews, essays on controversial topics, and personal interviews. Notably, the datasets cover the interplay between the two venues of deception: interactive/non-interactive and elicited/not elicited (Fitzpatrick et al., 2015). The online hotel reviews and essays represent non-interactive, elicited (through Amazon's Mechanical Turk) deception. Since the tasks are non-interactive, they do not necessarily require immediate responses. However, the two datasets differ because hotel reviews are about a service/product, while essays are introspective, personal beliefs. Meanwhile, the interview dataset is elicited deception that requires prompt responses in an interactive, conversational environment.

2.1.1 Deceptive Opinion Spam Corpus

Table 1: Corpus size and composition

Polarity	Truth type	
	Truthful	Deceptive
+	200	200
−	200	200

The Deceptive Opinion Spam Corpus was created by Ott et al. (2013) and consists of 400 truthful and 400 deceptive positive and negative reviews from TripAdvisor of 20 Chicago-area hotels. It serves as the "gold standard" for machine learning classifiers used to detect fake reviews because the truthfulness of each review is known. Table 2 shows a sample true and a sample deceptive review. Truthful reviews were obtained from automatically selecting valid reviews collected from TripAdvisor, while deceptive reviews were commissioned using Amazon's Mechanical Turk, where people get paid for completing online tasks.

Truth	Deceptive
I only stayed out with my boyfriend for one night, however enjoyed my stay. The staff was friendly, the room was nice and clean, the hallways and ballrooms etc were elegant. Room service was quick and had good options to choose from that actually tasted great. The staff was able to extend our check out time for an extra 1-2 hours without an extra charge to the room. Great location too! Walking distance from the Art Museum, Millennium Park, Grant Park (right across the street) and a quick cab ride to McCormick Place. If I were in the city again I would love to stay there again.	The Hilton in Chicago was awesome. The room was very clean and the hotel staff was very professional. One of the features I liked, was that in my room the internet access was wire and wireless, considering my laptop is not wireless, it help me out alot. Food was very good, quality was great. There was also a flat screen in my room...awesome. The hotel itself is locaated in the middle of alot of resturants with fin dinning. I also enjoyed the gym very much. Overall, I enjoyed myself, and I will stay again at the Hilton when I return to Chicago.

Table 2: Sample true and deceptive, positive online reviews from Ott et al. (2013).

2.1.2 Essays

Table 3: Corpus size and composition

Topic	Truth type	
	Truthful	Deceptive
Abortion	100	100
Death Penalty	98	98
Best Friend	98	98

Mihalcea and Strapparava (2009) created the Essays dataset, a labeled corpus containing true and deceptive opinions on controversial topics, such as abortion and the death penalty, as well as feelings about a best friend. This dataset was also created by using Amazon's Mechanical Turk.

Each essay topic contains approximately 100 essays per class containing statements like those reproduced from Mihalcea and Strapparava (2009) in Table 4. For the abortion and death penalty topics, the guidelines for the contributors were to

write at least 4-5 sentences about their true opinion on the topic as if they were preparing a speech for a debate, and then repeat the same process for their deceptive opinion. For the best friend topic, contributors were asked to write truthfully about the reasons they like their actual best friend, and deceptively about the reasons they like a person they could not stand.

True	Deceptive
Abortion	
I believe abortion is not an option. Once a life has been conceived, it is precious. No one has the right to decide to end it. Life begins at conception, because without conception, there is no life.	A woman has free will and free choice over what goes on in her body. If the child has not been born, it is under her control. Often the circumstances an unwanted child is born into are worse than death. The mother has the responsibility to choose the best course for her child.
Death penalty	
I stand against death penalty. It is pompous of anyone to think that they have the right to take life. No court of law can eliminate all possibilities of doubt. Also, some circumstances may have pushed a person to commit a crime that would otherwise merit severe punishment.	Death penalty is very important as a deterrent against crime. We live in a society, not as individuals. This imposes some restrictions on our actions. If a person doesn't adhere to these restrictions, he or she forfeits her life. Why should taxpayers' money be spent on feeding murderers?
Best friend	
I have been best friends with Jessica for about seven years now. She has always been there to help me out. She was even in the delivery room with me when I had my daughter. She was also one of the Bridesmaids in my wedding. She lives six hours away, but if we need each other we'll make the drive without even thinking.	I have been friends with Pam for almost four years now. She's the sweetest person I know. Whenever we need help she's always there to lend a hand. She always has a kind word to say and has a warm heart. She is my inspiration.

Table 4: Sample true and deceptive essays reproduced from Mihalcea & Strapparava (2009).

2.1.3 Deceptive Interview

Burgoon and Qin (2006) created the Deceptive Interview corpus by transcribing 122 verbal interview records where participants answered 12 interview questions, alternating between truthful and deceptive responses. Alternations occurred after every 3 questions with some participants beginning with a truthful block of questions and others beginning with a deceptive block. Table 5 includes a sample true and a sample deceptive answer to an interview question about the participant's educational background.

Truth	Deceptive
I have a bachelors of arts in education. I have an associates degree in accounting and computerized, eh um, bookkeeping and I have an artisans training in crafts. About eighteen years of formal school and about 45 years of practice. Oh yes, very much so. Um, not necessarily, I think a person who wants to be a teacher has to be very much dedicated, now more than ever. And as for accounting, that is just wisdom in these economic times. And I happen to be a creative fidget when it comes to crafts.	Well, I am a, I completed my masters degree in business administration. And I am hopefully going to be completing one for my doctorate, depending on time and money. In December of 1990. U of A. As I say that depends on money and the family situation. When I have time and money and work allows and everything else. Where did I complete that, I did that in '87, and I took some time off and went back. Here in Tucson.

Table 5: Sample true and deceptive responses to *'Please describe your educational background.'* from Burgoon & Qin (2006).

2.2 Related Work

Ultimately, automatic deception detection is a supervised learning problem. In supervised learning, a classifier, i.e., an algorithm that categorizes data into classes, is given labeled examples to learn from and then asked to classify new data it hasn't seen before. In other words, the classifier, or model, learns from its mistakes and tweaks its internal representation of how classification should operate to fix the predictions on its training data. After that, the model's performance is judged by its ability to correctly classify the unseen testing data.

Most early research on the benchmark dataset of Ott et al. (2011) has used *shallow* linguistic features, naive language properties that are simple to extract from

the text. These include n-grams over words (contiguous sequences of n words like *going+to*), part-of-speech (POS) tags (like ADJECTIVE), and *Linguistic Inquiry and Word Count* (LIWC) (Pennebaker, Booth, & Francis, 2007), which is a set of keyword lists for several semantic and psychological categories. Notably, Ott et al. (2011) demonstrated that an SVM classifier (a standard, widely used classification tool: Joachims, 2006) using only shallow features such as bigrams (2-grams) over words and the LIWC keyword lists was able to achieve 89.8% classification accuracy on the Deceptive Opinion Spam Corpus. While this is certainly impressive performance compared to the baseline figure of 50%, it's possible that greater performance improvements can be achieved by using *deeper* linguistic features. Deeper linguistic features can include aspects of language structure (such as the syntactic tree shown in Figure 1 below) and specific details, which get at underlying semantic components of the opinion. These can provide the model with a deeper understanding of what makes language deceptive or truthful.

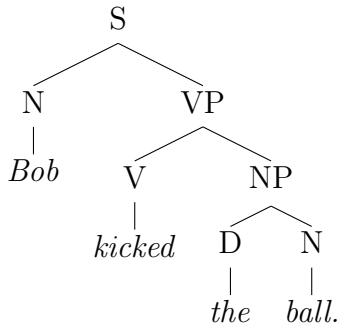

Figure 1: Syntactic parse tree, a type of deeper linguistic feature, for *Bob kicked the ball.*

Two more recent studies by S. Feng et al. (2012) and V. W. Feng and Hirst (2013) have successfully explored using deep features by building syntactic parse trees like the one in Figure 1. S. Feng et al. (2012) used both probabilistic context free grammar parse trees, a way of representing the tree structure of a sentence that incorporates how frequently different structures are used, and shallow features like POS-tags and word n-grams, to attain 91.2% accuracy on the Deceptive Opinion Spam Corpus dataset and 85% accuracy on the Essays dataset. In V. W. Feng and Hirst (2013), profile compatibility—a measure of how close an online review is to the other reviews for the same business or service—is added to the previous model of S. Feng et al. (2012). For each business's profile, different details (called "aspects" by V. W. Feng and Hirst (2013)) are automatically extracted. "Distinct" details

refer to proper noun phrases (e.g., *the Hilton*); "general" details refer to all other noun phrases (e.g., *the pool*, *the breakfast*). Both detail types are automatically extracted and clustered together to consolidate synonyms and semantically related terms (e.g., *price* and *rate* for a hotel). This compatibility process is intended to expose contradictions between a given review and a hotel's collective profile of other reviews. This latest model with a reimplemented baseline from other models is able to attain best performance of 91.3% accuracy on the Deceptive Opinion Spam Corpus dataset. This unfortunately is only a very small increase in performance over the 91.2% achieved by S. Feng et al. (2012).

3 Approach

3.1 Feature Engineering

Initially, we thought about clustering specific details into categories and creating features derived from these categories. For instance, V. W. Feng and Hirst (2013) perform hierarchical agglomerative clustering on a similarity metric derived from the WordNet ontology (Miller, 1995). We attempted to perform the same type of clustering using WordNet similarity metrics and GloVe word embeddings (Pennington et al., 2014). However, we found that evaluating the quality of the categories is challenging because the similarity of the details is difficult to quantify in a metric. Ideally, we would want an expert to hand-label some of the details as belonging to the same category so that we gauge the accuracy this way. Unfortunately, this hand-labeling is non-trivial and would need to be performed for each new domain introduced. Not to mention, it is unclear how many categories of specific details are ideal for each domain.

Instead, we opted for structural heuristics that use linguistic knowledge in order to capture various ways specific details surface in language. We engineered these features by manually examining numerous documents from each domain. The following features are extracted from each document after using a linguistic parser:

- **Prepositional Phrases**: Both the normalized number of prepositional phrase modifiers (the number of PP modifiers / the total number of phrases in the review) in the document and the average length of the PP modifiers.

- **Adjective Phrases**: Both the normalized number of adjective phrase modifiers (the number of ADJP modifiers / the total number of phrases in the review) in the document and the average length of the ADJP modifiers.

- **Numbers**: The total amount of numbers that occur in the document (i.e., words tagged with CD), normalized by the number of phrases in the document.
- **Proper Nouns**: The number of proper nouns (i.e., words tagged with NNP and NNPS) in a document that also appear in a dictionary, normalized by the total number of phrases in the document.
- **Consecutive Nouns**: The number of pairs of nouns (i.e., nouns that occur next to each other) in the document, normalized by the number of phrases in the document.

Intuitively, prepositional and adjectival phrase features aim to capture specific details such as mentioning "problems with the room's air conditioner" in a hotel review or spatial details about a hotel's location being "close to all the museums and theaters." The presence of proper nouns and consecutive nouns capture mentions of nearby places or amenities at a hotel, such as "Navy Pier", "Starbucks coffee", or an "airport shuttle."

3.2 Feature Analysis

After designing our features, we perform a sanity check to verify that the features are hallmarks of deception. To do so, we use a two-sample, independent t-test on the features to see if there is a significant difference between the feature values of truthful and deceptive document sets.

In Table 6, we see that, in total, 6 out of 15 deceptive features are statistically significant from their truthful counterparts. The combined modifying phrase count is robust across the domains of opinion spam and controversial essays. The only statistically significant feature in the interview dataset is the consecutive nouns.

4 Baselines and Future Work

4.1 Baseline Results

In order to evaluate the performance of our new feature set in the classification task, we must first test how a baseline model performs. We choose to implement the simple n-gram baseline used in Ott et al. (2013) for each dataset. We use an SVM classifier (a standard, widely used classification tool: Joachims, 2006) and transform our features using *term frequency-inverse document frequency* (*tf-idf*) from a bag-of-words representation. We use five-fold, nested cross-validation and report averaged

Table 6: Results from a two-sample, independent t-test of extracted features. Prepositional and adjective modifying phrases are combined into a single modifying phrase feature. The mean and standard deviation are reported for each truth type. Statistically significant results ($p \leq 0.05$) have their p-values bolded.

Feature	Op. Spam		Essays		Interview	
	T	D	T	D	T	D
mod. phrase count	1.7, 1.1	1.6, 0.7	1.6, 0.8	1.3, 0.9	1.1, 0.7	1.0, 0.7
	$p =$ **0.018**		$p =$ **0.0008**		$p = 0.12$	
mod. phrase length	3.1, 1.8	2.8, 1.8	2.4, 1.9	2.2, 1.9	2.1, 1.6	2.0, 1.5
	$p =$ **0.0002**		$p = 0.16$		$p = 0.20$	
numbers	0.2, 0.3	0.1, 0.2	0.1, 0.2	0.0, 0.1	0.1, 0.2	0.1, 0.2
	$p =$ **3.1e-23**		$p =$ **3.4e-7**		$p = 0.07$	
proper nouns	0.5, 0.6	0.5, 0.5	0.1, 0.4	0.2, 0.3	0.2, 0.3	0.2, 0.4
	$p = 0.73$		$p = 0.45$		$p = 0.80$	
consecutive nouns	0.5, 0.5	0.6, 0.4	0.3, 0.5	0.2, 0.4	0.2, 0.3	0.1, 0.2
	$p = 0.059$		$p = 0.07$		$p =$ **0.048**	

results of precision, recall, F_1 and accuracy over the folds in Tables 7, 8, and 9. We use the *scikit-learn* implementation for these tasks (Pedregosa et al., 2011).

4.1.1 Deceptive Opinion Spam Corpus

Table 7: Baseline results.

Feature	Polarity	Results			
		P	R	F_1	A
uni + bigram	+	88.7	88.6	88.6	88.6
	−	86.5	86.5	86.5	86.5

In Table 7, results are reported by training and testing on the same polarity (i.e., positive or negative reviews only). The *tf-idf* parameters (*norm*=$\ell 1$, *max doc. frequency*=0.15) and SVM's penalty parameter (C=175) were determined individually for each polarity through nested cross-validation.

4.1.2 Essays

Table 8: Baseline results.

Feature	Topic	Results P	R	F_1	A
uni + bigram	Abortion	69.1	67.5	66.8	67.5
	Death Penalty	59.7	59.7	59.6	59.8
	Best Friend	77.1	76.5	76.4	76.4

In Table 8, results are reported by training and testing only on the same essay topic. The *tf-idf* parameters (*norm*=$\ell 2$, *max doc. frequency*=0.2) and SVM's penalty parameter (C=0.001) were determined individually for each topic through nested cross-validation.

4.1.3 Deceptive Interview

Table 9: Baseline results.

Feature	Results P	R	F_1	A
uni + bigram	65.3	65.3	65.3	65.3

In Table 9, results are reported by training and testing on the entire dataset. training and testing on individual questions was attempted, but led to poor, near-chance results because of the sparsity of the data. The *tf-idf* parameters (*norm*=$\ell 1$, *max doc. frequency*=0.3) and SVM's penalty parameter (C=100) were determined through nested cross-validation.

4.2 Future Experiments

We plan to add our features to the existing n-gram baseline to test if they improve classification results. Then, we will examine the weights of our features to see if the SVM classifier considers them important. Further analysis will investigate if our features carry relevant information that no other feature contains (i.e., they are linearly independent from each other, and other features).

Additionally, we intend to perform many different training and testing procedures to understand the robustness of our features. For instance, we want to investigate

training on one dataset and testing on the others (both together and separately). Ultimately, we aim to answer the following questions: Will the inter-domain results be significantly better with our new features? How does the performance compare when training and testing on different combinations of datasets?

References

Burgoon, J. K., & Qin, T. (2006). The dynamic nature of deceptive verbal communication. *Journal of Language and Social Psychology*, *25*(1), 76–96.

Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (pp. 171–175).

Feng, V. W., & Hirst, G. (2013, October). Detecting deceptive opinions with profile compatibility. *International Joint Conference on Natural Language Processing*, 338-346.

Fitzpatrick, E., Bachenko, J., & Fornaciari, T. (2015). Automatic detection of verbal deception. *Synthesis Lectures on Human Language Technologies*, *8*(3), 1–119.

Joachims, T. (2006). Training linear svms in linear time. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (p. 217-226). New York, NY, USA: ACM.

Mihalcea, R., & Strapparava, C. (2009). The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In *Proceedings of the Association for Computational Linguistics.* Singapore: ACL.

Miller, G. A. (1995, November). Wordnet: A lexical database for english. *Commun. ACM*, *38*(11), 39–41.

Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative Deceptive Opinion Spam. In *HLT-NAACL* (pp. 497–501).

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 309–319).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennebaker, J., Booth, W., & Francis, M. (2007). *Linguistic inquiry and word count: Liwc.* Austin, TX: LIWC.net.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Emnlp* (Vol. 14, pp. 1532–1543).