New Methods for Detecting Deceptive Product Reviews

Andrea M. Kahn

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2015

Committee:

Yejin Choi, Chair

Fei Xia

Program Authorized to Offer Degree:
Department of Linguistics - Computational Linguistics

University of Washington

Abstract

New Methods for Detecting Deceptive Product Reviews

Andrea M. Kahn

Chair of the Supervisory Committee:
Assistant Professor Yejin Choi
Computer Science & Engineering

With the explosion of online shopping sites, there has been a proliferation of businesses offering to post positive product reviews in exchange for payment. The presence of these deceptive reviews transforms a product's online reviews from a source of candid customer feedback into a forum for surreptitious advertising. Given this trend, there is a demand for effective methods of detecting fake reviews and/or products for which reviews have been purchased. These tools could enable companies to identify and crack down on reviewers and vendors engaging in these deceptive practices.

In this work, we present new statistical methods for the detection of deceptive product reviews, focusing specifically on the detection of products for which a high percentage of the positive reviews are fake. Using a new hand-built corpus of online product reviews, we show that there is a correlation between textual features of a product's 5-star reviews and product metadata features that have been demonstrated to suggest the presence of deceptive product reviews, such as star rating distribution. Drawing from the literature on advertising language, the literature on deception, and a series of human performance experiments, we then propose a model that makes use of discourse structure to classify individual reviews as suspicious or trustworthy.

While there have been numerous studies in fake review detection, there has been relatively little work in identifying deception on the product level. Our work is also novel in that it

draws connections between fake review detection and advertising, a domain in which there has been little computational linguistics research. Furthermore, our work contrasts with much of the previous work in fake review detection in that we explore methods for developing a gold standard in a corpus of voluntarily posted product reviews, as opposed to soliciting deceptive reviews for the purpose of our research.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express her sincere appreciation to the University of Washington, the departments of Linguistics and Computer Science & Engineering, and her advisors Professor Yejin Choi and Professor Fei Xia. Thanks also go to Jun Seok Kang for assistance with web crawling, and to Professor Gina-Anne Levow for additional feedback on the project. The author was supported in part by a 2014-2015 Selected Professions Fellowship from the American Association of University Women (AAUW).

Chapter 1

INTRODUCTION

> *I was looking at my wife recently while we were taking in some Law & Order on the TV and she caught me staring at her. I noticed that her face seemed more radiant than usual and that her hair seemed healthy and vibrant as well. I was wondering if she was pregnant or something, so I asked. She was astonished that I even noticed, (we've been married forever) so she started telling me about this great natural oil she found on Amazon. It's called Argan oil and it is 100% organic and cold-pressed. She really liked that it is odorless because other products she has used can be really strong. Anyway, she said she puts it on her face about once a week and in her hair once a month. What was most important to her was that I NOTICED! This was money well spent in my opinion and I know she has probably told all of her friends about it. From a guy's point of view, the price is right and she looks better than ever.*
>
> —SonnyO, Amazon.com

Few people would disagree that there is something suspicious about this review. The challenge lies in pinpointing the characteristics that contribute to this suspiciousness. Is it the words the reviewer uses? The structure of the sentences? The structure of the storyline, or the existence of a storyline at all? Finally, does our own perception of the review's suspiciousness actually align with an intent to deceive on the part of the author?

In this work, we explore two contrasting but related questions: 1) Given a set of positive reviews for a particular product, can we determine whether the seller is likely to have written

and/or purchased some or all of the reviews?, and 2) Can we effectively characterize the language in product reviews that tends to arouse suspicion? While these two questions are related, they examine two different effects, since not all suspicious reviews are fake and not all fake reviews will arouse suspicion.

With the explosion of online shopping sites, there has been a proliferation of businesses offering to post positive product reviews in exchange for payment, a practice known as "astroturfing" [1] [2]. Many consumers rely on product reviews to make informed decisions about a purchase, and the presence of these deceptive reviews dilutes the candid customer feedback consumers expect with surreptitious advertising. Effective methods of detecting these deceptive reviews and/or products for which reviews have been purchased could enable companies to identify and crack down on reviewers and vendors engaging in these deceptive practices, improving customers' online shopping experience. Recently, large online retailers such as Amazon.com have invested increasing resources in developing machine-learning solutions for combating fake reviews [3].

In addition to the existence of companies that sell positive reviews, there appears to be a trend of sellers offering customers products for free or at reduced cost if they promise to post a review of the product. While many of the resulting reviews disclose that "[the customer] received this product in exchange for an honest review," there is reason to believe that there exist additional reviews written under these conditions that do not include this disclaimer. While customers who include this disclaimer claim not to be biased by the receipt of a free product and their promise to post a review in exchange, we have reason to believe otherwise: 92.6% of the reviews we collected that contain this or any of a set of related statements are 4- or 5-star, while only 80.8% of all reviews collected are 4- or 5-star. The presence of these reviews further highlights the transformation of online review boards into an advertising

---

[1] http://www.forbes.com/sites/retailwire/2015/04/13/amazon-lawsuit-takes-on-fake-reviewers/

[2] http://www.cbsnews.com/news/yelp-sues-companies-promising-positive-reviews/

[3] http://www.theguardian.com/technology/2015/jun/22/amazon-ai-fake-reviews-star-ratings-astroturfing

forum and supports the need for methods of detecting intentionally deceptive or otherwise biased reviews.

In this work, we present new statistical methods for the detection of deceptive product reviews, using a new hand-built corpus of Amazon.com product reviews. Through a series of binary classification experiments and correlation tests, we show that there is a correlation between textual features of a product's 5-star reviews and product metadata features that have been demonstrated to suggest the presence of deceptive product reviews, such as star rating distribution (Feng et al., 2012b). Drawing from the literature on advertising language, the literature on deception, and a series of human performance experiments, we then describe a model that makes use of discourse structure to classify individual reviews as suspicious or trustworthy.

Online shopping sites such as Amazon.com provide a rich corpus of text for analysis. However, a challenge associated with using real-life reviews for this type of research is that we have no reliable gold standard: We do not know for certain which reviews are fake or which products contain fake reviews. Nor do we want to label products manually, both because of the cost associated with doing this for a large dataset and because our judgments will be influenced in part by the very same textual features of reviews that we wish to use for classification. Thus, for the product-level experiments described in this work, we choose to use metadata features that we believe to be associated with deceptive products to automatically label products as deceptive or truthful. For the review-level experiments, we choose to focus on the detection of reviews that humans perceive as suspicious or trustworthy. While the categories that result from this distinction may deviate from the actual categories of real and fake reviews, automatically flagging reviews as suspicious or trustworthy is far more efficient than manually vetting reviews and/or products for suspicious activity.

This research draws from the related fields of deception detection, computational stylometry, and the linguistic analysis of advertising language. In turn, it provides insights into the relative effectiveness of approaches from each field for detecting deception in the product review domain.

Much of our work is motivated by distributional trends presented in Feng et al. (2012b). One major difference between our work and theirs is the fact that we use all of a product's reviews for classification, rather than using only the reviews of one-time reviewers for deceptive products and only the reviews of repeat reviewers for trustworthy products. In addition, we explore correlations between a range of textual features and star-rating skewedness, and we compare the effectiveness and transferability of classifiers trained on different categories of products.

While there have been numerous studies in fake review detection, there has been relatively little work in identifying deception on the product level (exceptions include work done by Feng et al. (2012b) and Li et al. (2013b)). Our work is also novel in that it draws connections between fake review detection and the language of advertising, a domain in which there has been little computational linguistics research. Furthermore, our work contrasts with much of the previous work in fake review detection in that we explore methods for developing a gold standard in a corpus of voluntarily posted product reviews, as opposed to soliciting deceptive reviews for the purpose of our research.

Chapter 2
LITERATURE REVIEW

Our research on fake review detection draws from the related fields of deception detection, computational stylometry, and the linguistic analysis of advertising language.

2.1 Deception Detection

Methods for identifying deceptive language are important for a variety of applications. In addition to facilitating the detection of online opinion fraud such as fake reviews, effective deception detection methods could have beneficial applications in fields such as politics, law enforcement, and academic publishing.

Studies of deception have sought to identify linguistic elements of deceptive speech and writing that differentiate it from truthful communication (Vrij, 2008; Dilmon, 2009). Vrij et al. (2009) found that those telling fabricated stories struggle with spatial representation. Ott et al. (2011) found that those who engage in deception tend to under-utilize personal pronouns, either to distance themselves from their lies or because they have not actually experienced the events they recount (Newman et al., 2003; Zhou et al., 2004). Pennebaker (2011) found similar distancing features in self-deceptive writing, including a lack of personal pronouns, a lack of emotional language, and a high density of discrepancy verbs like "would," "should," and "could" that create psychological distance. Zhou et al. (2008) present statistical language models that utilize n-gram features for online deception detection.

Deception studies in general suffer from a lack of a reliable gold standard. Deception researchers have dealt with this in a variety of ways. In their survey of deception studies, Gokhman et al. (2012) distinguish between sanctioned and unsanctioned deception. In studies that involve sanctioned deception, participants are instructed to lie. Examples include

studies by Newman et al. (2003) and Mihalcea and Strapparava (2009) in which participants were asked to lie about their beliefs regarding controversial topics. While this technique gives researchers greater control over the experiment and confidence in the gold standard, these experiments suffer from the limitation that participants are not engaging in "real" deception; individuals who have been instructed to lie may use language differently than those who lie of their own volition. In contrast, research in unsanctioned deception, such as an experiment conducted by Hancock et al. (2004) in which participants were asked to document their own acts of deception in a journal, rely on participants to truthfully report their own deception.

2.2 Computational Stylometry

The field of computational stylometry seeks to characterize text in terms of writing style. "Style" can encompass shallow lexico-syntactic features as well as deep syntactic elements such as parse-tree features. Subtasks of stylometric analysis include authorship attribution and text genre classification, as well as a number of domain-specific applications. Ashok et al. (2013) successfully apply stylometric analysis to predict the success of novels. Harpalani et al. (2011) use stylometric analysis to detect vandalism in Wikipedia edits.

Menon and Choi (2011) show that part-of-speech (POS) sequences and function word distributions can be used for cross-domain authorship attribution. We make use of these observations in our work, relying on the perplexity of the text of a product's reviews based on a POS language model as an indicator of whether a single author has created multiple reviews, a common deceptive practice in the online review space.

Rayson et al. (2001) found distinct differences in POS distributions across text genres, namely between informative and imaginative writing. Ott et al. (2011) found parallels between the POS distribution of deceptive reviews and the POS distribution of imaginative writing described by Rayson et al. (2001). These findings support the approach of using POS distributional features to detect deceptive reviews that we take in this work.

2.3 Fake Review Detection

Deceptive opinion spam is a growing problem in online review communities, and as a result, detecting deceptive opinion spam has become a topic of increasing interest to natural language processing and machine learning researchers. Ott et al. (2012) studied the rate of deceptive reviews on six popular online review sites in the travel and restaurant domain, finding that the relative growth rates of deception in each of these communities is related to the relative cost of deception in each community; practices such as filtering reviews written by first-time reviewers tend to quell the rate of deception. Automatic detection and filtering of deceptive reviews is particularly useful because studies show that humans are not able to detect deceptive reviews reliably (Ott et al., 2011).

Approaches to automatically detecting deceptive opinion spam can utilize purely textual features of reviews, metadata features of reviews and reviewers, or some combination of the two. Mukherjee et al. (2012) present approaches for spotting groups of deceptive opinion spammers in online review communities. They utilize frequent itemset mining to identify candidate groups of collaborative spammers, and then use metadata features such as the distance in time between the dates on which reviewers posted, deviation in star rating between a group's reviews and other reviews of the same products, and whether or not a group's reviews were among the first reviews posted for a particular product, to determine whether or not these groups are indeed perpetrators of collaborative opinion spamming. They find labeling groups of deceptive reviewers to be considerably easier than labeling individual spam reviewers.

An alternative approach to using metadata features to detect deceptive reviews or reviewers is treating fake review detection as a text categorization problem. Mihalcea and Strapparava (2009) present early work in detecting deception in written text, using word-class features. Feng et al. (2012a) build on work making use of lexical and shallow syntactic features for deception detection, showing that incorporating deep syntactic features results in improved classification accuracy on existing deception datasets. Li et al. (2014) build a

dataset for characterizing solicited deceptive reviews across three domains: hotels, restaurants, and doctors. Additionally, they contrast crowdsourced deceptive reviews written by Mechanical Turkers with deceptive reviews solicited from domain experts. They use word unigram, POS, and LIWC (Pennebaker et al., 2001) unigram features to classify reviews as deceptive or truthful, discovering that there are generalizable similarities in the distribution of these features in deceptive reviews across domains.

Li et al. (2013a) propose a topic modeling approach to fake review detection. This approach outperforms unigram and bigram baselines on Ott et al. (2011)'s hotel review dataset. Li et al. (2013b) use a semi-supervised manifold ranking algorithm and topic modeling features to identify offerings that contain a high density of deceptive reviews.

Like other areas in deception research, online opinion spam detection is a challenging domain because of the lack of a reliable gold standard: While one can derive heuristics for determining the trustworthiness of a review based on review or reviewer metadata (e.g., a large number of 2-3 word 5-star reviews written over a short period of time are likely to signify suspicious activity; the reviews of reviewers who have only posted one review, post only 2-3 word 5-star reviews, or have a generic username followed by a long string of digits should be regarded with suspicion), it is usually impossible to say for certain whether or not a review is fake. One could even argue that the distinction between "real" and "fake" reviews is a blurry one; recently, there appears to be a trend of companies offering potential customers complimentary or discounted products in exchange for the promise of a review. While many of these reviews include a disclaimer stating that the content is the reviewer's unbiased opinion, these reviews are more positive than other reviews on average; 92.6% of the reviews in our dataset that contain this type of disclaimer are 4- or 5-star, while only 80.8% of all reviews in our dataset are 4- or 5-star. In a study of Amazon's Vine program, through which reviewers are provided with free products in exchange for posting reviews, Puranam et al. (2014) find linguistic differences in participant's reviews after enrollment in the program, including longer reviews, more complex sentence structure, more positive emotion, and more physical and personal descriptions of the product. This so-called "enrollment effect" most

likely extends beyond official programs such as Amazon Vine to customers provided with promotional products for review from companies themselves.

Many studies have alleviated the problem of a lack of a gold standard by crowdsourcing fake reviews. An example is the work of Ott et al. (2011), in which a corpus of deceptive hotel reviews was built using Amazon Mechanical Turk (MTurk) to solicit deceptive reviews. The problem with this approach parallels the problem with sanctioned deception approaches described above: These reviews are not written under the same circumstances or by the same authors as the fake reviews this research attempts to characterize, and one can speculate that these differences could have noticeable effects on the content and style of these reviews. Indeed, Li et al. (2014) note important differences between deceptive reviews written by Mechanical Turkers, which tend to exhibit POS features typical of imaginative writing, with those solicited from domain experts, which tend to exhibit POS features of informative writing. Thus, there is a tradeoff in the field of fake review detection between using "fake" fake reviews and having a reliable ground truth on the one hand, and using "real" fake reviews but having no reliable gold standard on the other.

An alternative approach to the problem of developing a gold standard is heuristic labeling. Jindal and Liu (2008) propose using duplicate reviews as suspicious examples in hand-curated corpora for fake review detection. This method may be out-dated, as many review sites now have methods for detecting and deleting duplicate reviews. Wu et al. (2010) propose a distortion criterion for evaluating the effectiveness of fake review detection methods, using a hand-built corpus of TripAdvisor hotel reviews. They claim that deceptive reviews will distort the popularity ranking of hotels more than other reviews, and that a review's deceptiveness can be measured by the change in the hotel's popularity ranking when the review is removed. Their evaluation of this approach is itself based on a heuristic: namely, the assumption that reviews posted by one-time reviewers are more likely to be deceptive.

Feng et al. (2012b) propose that there is a "natural" distribution of review star ratings, and that businesses that hire writers to generate positive reviews will have distinct star rating distributions from those that do not engage in this deceptive practice. They also show

that one-time reviewers, who are generally accepted to be more likely to be spammers, are also more likely to leave 5- or 1-star reviews than writers of multiple reviews. Thus, particularly skewed star-rating distributions can be used as a labeling heuristic in the detection of businesses engaging in deceptive activity in the online review space. This is the primary heuristic that we use for labeling in this work. One major difference between our work and that of Feng et al. (2012b) is the fact that we use all of a product's reviews for classification, rather than using only the reviews of one-time reviewers for deceptive products and only the reviews of multi-time reviewers for trustworthy products.

While fake review detection draws approaches from the field of deception detection, there is evidence that domain-specific approaches are necessary to achieve optimal performance, as is the case with many natural language processing tasks. In particular, taking into account context is very important. Ott et al. (2011) found that a model trained on word unigrams and bigrams outperformed a model trained solely on traditional deception cues. Ott et al. (2011) also observed many consistent elements of deceptive product reviews that were the opposite of traditional deception cues: For example, deceptive reviews tended to contain a high concentration of first-person singular pronouns, whereas Newman et al. (2003) found deceptive text to contain a lower concentration of personal pronouns, which they attribute to the deceivers' desire to distance themselves from their own deception. These contrasts between deception cues in the online review space and traditional deception cues highlight the need for alternative approaches for deception detection in the domain of online reviews.

While we focus in this work on detecting deceptive positive reviews posted by businesses to artificially inflate their ranking on online review sites, there exists a parallel problem of detecting deceptive negative reviews posted by competitors. Ott et al. (2013) built a dataset of deceptive negative reviews using the same crowdsourcing approach taken in Ott et al. (2011). They found deceptive negative reviews to contain POS distributions characteristic of imaginative writing, fewer spatial relationships, and more exaggerated sentiment. These findings are consistent with observations by Ott et al. (2011) regarding deceptive positive reviews. In contrast with the findings of Ott et al. (2011), however, Ott et al. (2013) they

observe less of an increase in personal pronoun usage over truthful reviews.

2.4 Linguistic Analysis of Advertising

One could characterize deceptive positive reviews as a form of surreptitous advertising. Thus, we turned to the literature on advertising for deriving features for detecting deceptive product reviews. In many cases, such as that of personal pronoun use, characteristics of advertising text contrast with traditional deception markers.

According to the advertising literature (and confirmed by our own observations), advertising language is characterized by an unusually high concentration of imperative sentences, ostensibly with the intent of inciting the reader to action (Leech, 1966; Vestergaard and Schrøder, 1985). Leech (1966) also observes that advertising language commonly contains a relatively high concentration of imperative sentences, often appearing to serve the purpose of singling out and forming a connection with the ad's target audience. Cook (2001) also observe that advertisements make heavy use of phonological and syntactic parallelism.

Cook (2001) and Fuertes-Olivera et al. (2001) observe a high density of personal pronouns in advertising, hypothesizing that their function is to establish a connection between the advertiser and the audience. This contrasts with findings in the field of deception detection, but aligns with observations made by Ott et al. (2011) regarding fake review detection, suggesting that deceptive reviews may share more elements with advertisements than they do with deceptive language in other domains.

To augment the observations from the advertising literature described above, we perused online advertisement corpora such as `www.vintageadbrowser.com` to come up with a few of our own. One observation we made was that advertisements often use null comparatives: comparative words without an object. Examples include a vintage Lucky Strike ad that asserts "Luckies Taste Better! Cleaner, Fresher, Smoother!", and a Gillette razor blade ad that claims "Gillette blades are harder, stronger, keener, longer" (Gieszinger, 2001). We utilize these observations in our construction of hand-crafted features for the detection of deceptive reviews and products.

Chapter 3
METHODOLOGY

The experiments described below fall into three major categories: correlation tests, classification experiments, and human performance experiments. A further distinction can be drawn between experiments that focus on detecting suspicious products and those that focus on detecting suspicious reviews.

3.1 Data

We collected over 1 million reviews for over 16,000 products from the online retail site Amazon.com between April and June of 2015. The resulting datasets were used in the correlation and classification experiments described in sections 3.2 and 3.3.

Reviews and product metadata were gathered using screen-scraping techniques. An initial script gathered URLs for the 5,000 top-listed products in each of 10 categories. While it is not clear how Amazon decides how to order their products, the top-listed products appeared to have more reviews on average. A subsequent script crawled the product pages for product metadata and all reviews. (The final datasets contain fewer than 5,000 products because many products do not have any reviews, and some categories had fewer than 5,000 products.)

We removed all products that did not have at least one review. For each of the 10 categories, we randomly selected 20% of the remaining products to be added to a held-out test set.

A summary of the datasets is given in tables 3.1 and 3.2. Experiments were performed on datasets for individual categories of products, as well as on the complete set combining products from all categories. The advantage of performing experiments on an individual category is that it protects against the classifier learning to detect category boundaries,

which may also correlate with the metadata features we use to label the data. For example, vitamins may have higher star ratings on average than hair products; a classifier that uses word unigram features can easily learn to distinguish between hair products and vitamins, and therefore may appear to be performing very well if we have used the average star rating to label the data.

The numbers for the combined dataset differ slightly from the sum of the columns because a number of products (approximately 200) were found in more than one dataset. If a duplicate product was found in the training set for one category and the test set for another, it was removed from the test set in the combined dataset, so that models trained on training instances from a particular category and tested on test instances for all categories would not be tested on instances they had already seen.

Because we were particularly interested in analyzing the linguistic features of fairly long reviews, we chose to focus on categories of products around which customers tend to form intricate personal narratives, such as vitamins, skin and hair products, and lingerie. We were also less interested in detecting counterfeit versions of brand-name products than we were in detecting products for which a large number of reviews had been written by the seller or others with ulterior motives (e.g., hired professional writers); therefore, we focused on categories of products that could be produced in small batches at relatively low cost by a small company (e.g., skin and hair products), as opposed to products for which counterfeit versions are the primary concern (e.g., electronics or jeans). The clothing categories were selected for contrast, based on the hypothesis that these would contain fewer counterfeit items, and, as a result, fewer fake reviews.

3.2 Feature Correlation Tests

In an effort to discover detectable textual and metadata features tending to correlate with fake reviews, we performed a number of correlation tests. For each test, we calculated the correlation between a pair of features using Pearson's correlation coefficient, given by the following equation:

Product Category	**Number of Products**	**Number of Reviews**	**Avg. Reviews per Product**	**Avg. Star Rating**
Bath & Body	1,791	107,388	59.96	4.35
Cough	2,344	34,102	14.55	4.40
Cleaning	1,455	107,682	74.01	4.29
Fragrance	1,603	103,289	64.43	4.41
Hair	841	147,585	175.49	4.23
Herbal Supplements	2,148	104,425	48.61	4.36
Women's Lingerie	666	108,499	162.91	4.11
Skin	937	147,156	157.05	4.33
Men's Underwear	2,688	93,793	34.89	4.20
Vitamins	2,531	108,609	42.91	4.44
Combined	16,812	1,041,588	61.96	4.33

Table 3.1: Datasets Used for Correlation and Product Classification Experiments

Product Category	**Number of Products**	**Size of Training/Validation Set**	**Size of Test Set**
Bath & Body	1,791	1,433	358
Cleaning	1,455	1,164	291
Cough	2,344	1,876	468
Fragrance	1,603	1,283	320
Hair	841	673	168
Herbal Supplements	2,148	1,719	429
Women's Lingerie	666	533	133
Skin	937	750	187
Men's Underwear	2,688	2,151	537
Vitamins	2,531	2,025	506
Combined	16,812	13,502	3,310

Table 3.2: Data Used for Correlation and Product Classification Experiments: Training and Test Splits

$$\rho_{f_1,f_2} = \frac{cov(f_1, f_2)}{\sigma_{f_1}\sigma_{f_2}}$$

Tests fell into two categories: those for which product-level feature values were used, and those for which review-level feature values were used.

Since we did not know which reviews were fake (or which products had a high incidence of fake reviews), each pair consisted of an indicator feature that we believed to correlate with fake reviews, which served as a rough measure of the review's (or the product's) "suspiciousness." In the case of products, this was either the "star-rating skewedness," a calculation of the percentage of reviews that were 5-star (or 5- and 1-star), or the variance of the reviews' creations dates. In the case of reviews, this was either star-rating skewedness or the date on which the review was created. The choice of star-rating skewedness as an indicator feature was based on previous research suggesting that the star ratings of suspicious products tend to form a bimodal distribution (Feng et al., 2012b). The choice of the variance of the creation dates of a products' reviews was based on the hypothesis that products for whom a seller has purchased reviews tend to have a high density of reviews in a short period of time. Finally, the choice of the review creation date as an indicator feature was based on the hypothesis that the incidence of fake reviews was likely to increase over time.

In the classification experiments described in section 3.3, we only extracted textual features from a product's 5-star reviews. This decision was based on the intuition that the deceptive reviews we were interested in detecting (reviews solicited by the seller) were most likely to be 5-star reviews, as well as the fact that the labels for the classification experiments were generated using the review distribution and thus using the full distribution of reviews might be seen as "cheating." If a product had no 5-star reviews, it was not considered in this experiment.

A complete list of correlation tests performed and results is given in section 4.1.

3.2.1 Features

Table 3.3 describes the features that were explored in correlation tests. These features comprise standard n-gram features used in text classification as well as a number of hand-crafted features based on the advertising literature. Observations from the human performance experiments described in section 3.4 were also taken into account in feature engineering.

With the exception of the `skewedness` and `bimodal_skew` features, all product-level features were calculated for a product's 5-star reviews only. Similarly, in the review-level correlation tests, only 5-star reivews were used as data points.

Sentiment features were normalized counts of unigrams appearing in the sentiment lexicon provided in Hu and Liu (2004).

The feature `science_words` was calculated using a small lexicon of scientific and pseudo-scientific words that we created after reading a number of suspicious-sounding reviews that appeared to be attempting to establish trust in the reader by citing "facts" about product ingredients or relevant alleged scientific studies. Examples include *scientific*, *natural*, *organic*, *research*, *proven*, and *patent*.

The feature `comparatives` was generated by using NLTK (Bird et al., 2009) to label parts of speech in the text and counting the frequency of JJR tags, then normalizing by the number of tokens. To calculate the value for `null_comparatives`, we searched in a fixed window after JJR tags for an occurrence of the word *than*. We calculated the value for `imperatives` using a heuristic: We counted the number of sentences beginning with any of the 100 most common English verbs in their base form.

Our decision to use perplexity as a feature was based on the hypothesis that products with suspicious activity would be likely to have a large number of reviews written by the same person, which would be more linguistically homogeneous than reviews written by different authors.

FEATURE	LEVEL	DESCRIPTION
skewedness	product-	The percentage of the product's reviews that are 5-star
bimodal_skew	product-	The percentage of the product's reviews that are 5- or 1-star
creation_date	review-	The date on which the review was created
creation_month	review-	The month and year in which the review was created
creation_year	review-	The year in which the review was created
date_variance	product-	The variance of the creation dates of the product's reviews
pronouns	product-; review-	The percentage of words in the review(s) that are in a list of pronouns
science_words	product-; review-	The percentage of words in the review(s) that are in a list of "scientific" words we created
sentiment	product-; review-	The percentage of words in the review(s) that are sentiment words
pos_sentiment	product-; review-	The percentage of words in the review(s) that are positive sentiment words
neg_sentiment	product-; review-	The percentage of words in the review(s) that are negative sentiment words
imperatives	product-; review-	The count of imperatives (normalized by the number of sentences) in the review(s)
comparatives	product-; review-	The count of comparatives (normalized by the number of tokens) in the review
null_comparatives	product-; review-	The count of null comparatives (normalized by the number of tokens) in the review
name_mentions	product-; review-	The count of product, brand, and seller mentions (normalized by the number of sentences) in the review(s)

avg_word_len	product-; review-	The average length in characters of all words in the review(s)
avg_sent_len	product-; review-	The average length in words of all sentences in the review(s)
avg_review_len	product-; review-	The average length in words of the product's reviews, or the length of the review
word_perplexity	product-	The perplexity of a word-unigram language model of the product's reviews
pos_perplexity	product-	The perplexity of a POS-unigram language model of the product's reviews

Table 3.3: Features Explored in Correlation Tests

3.3 Text-Based Product Classification

We explored the effectiveness of classifying products using various textual features of their 5-star reviews, using metadata features to generate the gold-standard class labels. The goal of these experiments was to determine whether the 5-star reviews of a suspicious product exhibited generalizable differences from the 5-star reviews of a typical product. All classification experiments were performed with MALLET (McCallum, 2002) using a maximum entropy classifier and 4-fold cross-validation.

For these experiments, products were labeled as "suspicious" or "trustworthy" based on the percentage of their reviews that were 5-star, or the percentage of their reviews that were 5- or 1-star. Products for which the percentage was high were considered to be suspicious, while products for which the percentage was low were considered to be trustworthy. For each experiment, a minimum threshold for this percentage was selected for suspicious products, and a maximum threshold was selected for trustworthy products. The thresholds were always set such that each class would have the same number of products.

Because the gold-standard labels were based on the star rating distribution of products' reviews, textual features were extracted from a product's 5-star reviews only. This was done to prevent the model from classifying on the basis of the contrasting sentiment of reviews of different star ratings. If the classifier learned to predict star ratings, it would be able to predict labels that had themselves been generated from distributions of star ratings with high accuracy, but it would not provide insight into our reasearch question. Furthermore, deceptive reviews are most likely to be 5-star or 1-star (depending on whether they are solicited by the seller or written by a competitor), and we were primarily interested in characterizing deceptive positive reviews. During the labeling step, we discarded products that did not have at least one 5-star review.

During the training and tuning process, experiments were conducted using k-fold cross-validation on the full training/validation set. Parameters were tuned to achieve optimal results on the full training/validation set. During the final testing phase, models were trained on the training/validation set for each category of products and the training/validation set of the combined dataset, and tested on the held-out data from the same category as well as the held-out data from all categories (to explore the versatility of the category-specific models).

The results of these experiments are summarized in section 4.2.

3.3.1 Features

Table 3.4 describes the features that were explored in product classification experiments. Many of these features were those explored in the correlation tests described in section 3.2. We also used word and part-of-speech (POS) tag n-gram features, as well as sentiment words only and function words only. POS tags were generated using NLTK's POS tagger (Bird et al., 2009). We used the sentiment lexicon provided in Hu and Liu (2004) for sentiment features and the stopword list provided with the Lemur Project for function words.

FEATURE	DESCRIPTION
unigrams	Word unigram features, normalized by the total unigram count
bigrams	Word bigram word features, normalized by the total bigram count
trigrams	Word trigram features, normalized by the total trigram count
skipgrams	Skip-one word bigram features, normalized by the total trigram count
pos_unigrams	POS unigram features, normalized by the total unigram count
pos_bigrams	POS bigram features, normalized by the total bigram count
sentiment	Sentiment word features, normalized by the total word count
funct_words	Function word features, normalized by the total word count
avg_word_len	The average length in characters of all words in the review(s)
avg_review_len	The average length in words of the product's reviews, or the length of the review
pos_perplexity	The perplexity of a POS-unigram language model of the product's reviews

Table 3.4: Features Used in Product Classification Experiments

3.3.2 Model Parameters

Table 3.5 describes additional tuned parameters of the model. We used the training/validation set to discover optimal parameter settings for each category of products as well as for the full dataset.

3.4 Human Performance Experiments

We performed a series of Amazon Mechanical Turk (MTurk) experiments in which we asked Turkers to report their perceptions of whether or not a particular review was trustworthy.

PARAMETER	DESCRIPTION
`verified_only`	Only consider reviews marked 'Verified Purchase'
`stoplist`	Use a stopword list for word unigram features
`min_word_unigram_val`	Only keep word unigram features with this value or higher
`min_review_len`	Only consider reviews of at least N words
`max_num_reviews`	Consider at most N reviews; sort reviews by length and take the top N

Table 3.5: Model Parameters for Text-Based Product Classification

These experiments had three aims: 1) to evaluate the consistency of human perceptions of trustworthiness (i.e., through calculating inter-rater agreement); 2) to collect labels for subsequent review classification experiments described in section 3.5; and 3) to gather ideas for custom features for the classifiers described in sections 3.3 and 3.5.

Each Human Intelligence Task (HIT) consisted of a single review. Reviewers were presented with the following instructions:

> *When you are browsing an online shopping site, you expect that reviews have been written by customers who have purchased and used the products they are writing about. Sometimes, however, a positive review is posted by someone other than an ordinary customer—either by the seller of the product, or by a company hired to write positive reviews.*
>
> *Please read the positive review below and answer the following questions about it. Keep in mind that a trustworthy review is one that appears to be written by an actual customer, and a suspicious review is one that appears to be written by a seller or advertiser masquerading as a customer.*

Turkers were then asked to rate the review as "Very Trustworthy," "Somewhat Trustworthy," "Somewhat Suspicious," or "Very Suspicious." In addition, they were asked to select all sentences from a list of sentences from the review that struck them as suspicious, and all sentences that struck them as trustworthy. Finally, they were given an opportunity to explain their responses to the questions in a text box.

We performed the experiment for 48 reviews (3 batches of 16 reviews each). We used workers with the Masters qualification only and paid $0.16 per HIT.

We selected reviews from a variety of products across categories. In general, we selected reviews that were longer than average and that struck us as interesting, either due to a complex narrative or idiosyncratic language use. We tried to balance the number of reviews that we perceived as trustworthy and the number that we perceive as suspicious.

In chapter 4, we calculate inter-rater agreement and share interesting insights gathered from the open-ended questions.

3.5 Discourse-Based Review Classification

We conducted a small exploration of the effectiveness of review classification based on discourse modeling. The motivation behind these experiments was the idea that suspicious reviews tend to share a similar content structure. These experiments were limited and much remains to be explored in this domain; see chapter 7.

While we derived a metric for classifying products as likely to exhibit suspicious activity or not (namely, their star rating distribution), a similar metric does not exist for individual reviews. Therefore, in these experiments, we attempted to characterize the human perceptions of suspicion explored in section 3.4, as opposed to the actual presence of suspicious activity explored in sections 3.2 and 3.3.

3.5.1 Data

The dataset consisted of 42 reviews. Binary labels ("suspicious" or "trustworthy") for the reviews were chosen based on the majority label collected in the MTurk experiments described

in section 3.4.

3.5.2 Sentence Labeling

Sentences were labeled using the labeling scheme described in table 3.6. Each sentence was given exactly one label. Since multiple labels could theoretically apply to a single sentence, labels were considered in a hierarchical fashion; for example, if a sentence could be classified as `FACTUAL` (the first label in the list), that label was assigned, and no subsequent labels were considered. Labels were context-sensitive, meaning that the same sentence could be assigned a different label depending on the sentences that surrounded it.

Labels were chosen according to 1) their usefulness in modeling the types of content that comprise reviews, particularly content that contributes to readers' perception of the reviews as suspicious or trustworthy, and 2) the ease with which a classifier using n-gram features could be trained to recognize sentences belonging to these categories.

We manually labeled 441 sentences (131 `OPINION`, 49 `FACTUAL`, 158 `ANECDOTE`, 103 `HISTORY`), and experimented with training a sentence classifier using word and POS n-gram features. The experiment was conducted in MALLET, using a maximum entropy classifier and 4-fold cross-validation.

3.5.3 Review-Level Classification

Based on the results of the human performance experiments described in section 3.4, we hypothesize that reviews that arouse suspicion in humans exhibit a distinctive content structure that is unique from that of reviews humans find trustworthy. Previous work in discourse modeling utilizes sentence-level information to make document-level predictions; for example, Barzilay and Lee (2004) use content models to generate extractive summaries of documents, and McDonald et al. (2007) develop a joint model for classifying documents as exhibiting positive or negative sentiment based on sequences of automatically generated positive or negative sentence labels.

LABEL	DESCRIPTION	EXAMPLE
FACTUAL	Sentence relates impersonal, objective facts about the product (uses, ingredients, etc.).	*It's called Argan oil and it is 100% organic and cold-pressed.*
ANECDOTE	Sentence comprises a personal anecdote relating to use of the product.	*I noticed that her face seemed more radiant than usual and that her hair seemed healthy and vibrant as well.*
HISTORY	Sentence provides personal information about the reviewer, family members, etc., and/or describes their life or impressions prior to using the product.	*I'm 27 years old, and I've suffered from all sorts of skin problems since the age of 5.*
OPINION	Sentence expresses a subjective opinion about the product, the seller, or Amazon.	*I think the price is good for the quality of product you are getting, and how good you will look.*

Table 3.6: Sentence-Level Labels for Discourse-Based Classification

Taking these approaches as inspiration, we explored training a classifier to identify suspicious reviews using n-gram features generated from the sequence of sentence labels in a review. As an example, if a review's sentences were labeled ANECDOTE FACTUAL ANECDOTE ANECDOTE, bigram features would include: START_ANECDOTE, ANECDOTE_FACTUAL, FACTUAL_ANECDOTE, ANECDOTE_ANECDOTE, and ANECDOTE_STOP.

Though the ultimate goal would be to chain a sentence-level classifier before a review-level classifier to automatically generate sentence labels for the sentence-label features, we used manually generated sentence labels for these experiments due to the relatively low performance of the initial n-gram based sentence classifier.

Chapter 4

RESULTS

4.1 Feature Correlation Tests

To better characterize the language of deceptive product reviews, we calculated Pearson's correlation coefficient for various pairs of features y, x on the 5-star reviews in the training set for three unique categories of products, as well as the training set for all products. Since the reviews themselves were not labeled as real or fake, y was an indicator feature we believed to be correlated with the trustworthiness of the review; x was a feature (typically a textual feature) we wished to explore. This also helped us assess the informativeness of various textual features for the product classification experiments described in sections 3.3 and 4.2.

Tests fell into two categories: those for which product-level feature values were used, and those for which review-level feature values were used. Section 3.2 provides a more in-depth description of these experiments and their justification; table 3.3 gives a summary of the product- and review-level features explored. For the product-level experiments, all features are calculated at the product level; for the review-level experiments, all features are calculated at the review level. (For review-level experiments, "average review length" is simply the length of the review.)

Tables 4.1-4.8 summarize the results of the experiments. Boldface correlation values are significant at a p-value <0.01. Non-italicized correlation values are significant at a p-value <0.05. Italicized correlation values have a p-value >0.05 (i.e., are not significant).

4.2 Text-Based Product Classification

We explored the effectiveness of classifying products using various textual features of their 5-star reviews, using products' star-rating distribution and sliding minimum and maximum

	Date Variance	Avg. Word Length	Avg. Sentence Length	Avg. Review Length	Word Perplexity	POS Perplexity
Products: Star-Rating Skewedness (5- and 1-star)	*-0.0180*	**0.0791**	*-0.0187*	**-0.0309**	*0.0038*	**-0.1037**
Products: Star-Rating Skewedness (5-star)	**-0.0607**	**0.0990**	**-0.0450**	**-0.0781**	*0.0011*	**-0.1187**
Products: Review Creation Date Variance	N/A	**-0.1139**	**0.1781**	**0.1833**	*-0.0166*	**0.2741**
Reviews: Creation Date	N/A	**0.0849**	**-0.1302**	**-0.1422**	N/A	N/A
Reviews: Creation Month	N/A	**0.0848**	**-0.1302**	**-0.1422**	N/A	N/A
Reviews: Creation Year	N/A	**0.0832**	**-0.1270**	**-0.1378**	N/A	N/A

Table 4.1: Results of Feature Correlation Tests on Training Set, Non-Lexical Features (All Products)

	Sentiment	Positive Sentiment	Negative Sentiment	Pronouns	"Scientific" Words	Imperatives	Comparatives	Null Comparatives	Brand Names
Products: Star-Rating Skewedness (5- and 1-star)	**0.0540**	**0.0368**	**0.0831**	-0.0190	**0.0855**	**0.0265**	**-0.0415**	**-0.0404**	**-0.0503**
Products: Star-Rating Skewedness (5-star)	**0.0718**	**0.0565**	**0.0704**	**-0.0433**	**0.0914**	**0.0241**	**-0.0244**	-0.0251	**-0.0345**
Products: Review Creation Date Variance	**-0.1593**	**-0.1692**	**0.0757**	**0.0986**	*0.0124*	*-0.0032*	*-7E-05*	*0.0113*	*-0.0151*
Reviews: Creation Date	**0.2444**	**0.2491**	**-0.0244**	**-0.0812**	**-0.0065**	**0.0049**	**-0.0139**	**-0.0144**	**-0.0292**
Reviews: Creation Month	**0.2444**	**0.2491**	**-0.0244**	**-0.0812**	**-0.0065**	**0.0049**	**-0.0139**	**-0.0144**	**-0.0292**
Reviews: Creation Year	**0.2359**	**0.2404**	**-0.0235**	**-0.0765**	**-0.0060**	**0.0044**	**-0.0133**	**-0.0135**	**-0.0287**

Table 4.2: Results of Feature Correlation Tests on Training Set, Word-Class Features (All Products)

	Date Variance	Avg. Word Length	Avg. Sentence Length	Avg. Review Length	Word Perplexity	POS Perplexity
Products: Star-Rating Skewedness (5- and 1-star)	*-0.0303*	**0.1327**	**-0.2415**	**-0.2604**	*-0.0264*	**-0.1037**
Products: Star-Rating Skewedness (5-star)	*-0.0815*	**0.1750**	**-0.2758**	**-0.2824**	*-0.0475*	**-0.1187**
Products: Review Creation Date Variance	N/A	*-0.0191*	**0.2958**	**0.2974**	*0.0299*	0.2741
Reviews: Creation Date	N/A	**0.0283**	**-0.1417**	**-0.1644**	N/A	N/A
Reviews: Creation Month	N/A	**0.0282**	**-0.1417**	**-0.1644**	N/A	N/A
Reviews: Creation Year	N/A	**0.0250**	**-0.1409**	**-0.1629**	N/A	N/A

Table 4.3: Results of Feature Correlation Tests on Training Set, Non-Lexical Features (Skin Products)

	Sentiment	Positive Sentiment	Negative Sentiment	Pronouns	"Scientific" Words	Imperatives	Comparatives	Null Comparatives	Brand Names
Products: Star-Rating Skewedness (5- and 1-star)	**0.1746**	**0.1785**	*0.0104*	*-0.0362*	0.0261	-0.0320	-0.0901	**-0.1080**	**-0.0907**
Products: Star-Rating Skewedness (5-star)	**0.2121**	**0.2208**	*0.0015*	*-0.0637*	*0.0437*	*-0.0268*	-0.0778	**-0.1050**	-0.0565
Products: Review Creation Date Variance	**-0.2133**	**-0.2393**	*0.0319*	*-0.0091*	**0.1090**	-0.0498	**-0.0954**	-0.0775	0.0102
Reviews: Creation Date	**0.2446**	**0.2481**	**-0.0169**	**-0.0745**	**-0.0189**	**0.0137**	**-0.0135**	**-0.0148**	**-0.0247**
Reviews: Creation Month	**0.2446**	**0.2481**	**-0.0169**	**-0.0743**	**-0.0189**	**0.0137**	**-0.0135**	**-0.0148**	**-0.0246**
Reviews: Creation Year	**0.2370**	**0.2406**	**-0.0170**	**-0.0719**	**-0.0179**	**0.0134**	**-0.0120**	**-0.0127**	**-0.0233**

Table 4.4: Results of Feature Correlation Tests on Training Set, Word-Class Features (Skin Products)

	Date Variance	Avg. Word Length	Avg. Sentence Length	Avg. Review Length	Word Perplexity	POS Perplexity
Products: Star-Rating Skewedness (5- and 1-star)	**-0.0841**	**0.1236**	**-0.0785**	**-0.0944**	*-0.0187*	**-0.1092**
Products: Star-Rating Skewedness (5-star)	**-0.1099**	**0.1388**	**-0.0822**	**-0.0991**	*-0.0183*	**-0.1048**
Products: Review Creation Date Variance	N/A	**-0.0783**	**0.1534**	**0.0824**	*0.0178*	**0.1989**
Reviews: Creation Date	N/A	**0.0501**	**-0.0749**	-0.0086	N/A	N/A
Reviews: Creation Month	N/A	**0.0502**	**-0.0751**	-0.0088	N/A	N/A
Reviews: Creation Year	N/A	**0.0478**	**-0.0682**	*0.0023*	N/A	N/A

Table 4.5: Results of Feature Correlation Tests on Training Set, Non-Lexical Features (Vitamins)

	Sentiment	Positive Sentiment	Negative Sentiment	Pronouns	"Scientific" Words	Imperatives	Comparatives	Null Comparatives	Brand Names
Products: Star-Rating Skewedness (5- and 1-star)	**0.0674**	**0.0746**	*-0.0338*	**-0.0911**	*-0.0089*	*0.0375*	*-0.0231*	*-0.0120*	*0.0326*
Products: Star-Rating Skewedness (5-star)	0.0538	**0.0642**	-0.0453	**-0.1338**	*0.0192*	*0.0311*	*0.0007*	*0.0006*	*0.0347*
Products: Review Creation Date Variance	**-0.1316**	**-0.1448**	**0.0632**	*0.0309*	*-0.0046*	*0.0196*	*0.0049*	*0.0036*	*-0.0283*
Reviews: Creation Date	**0.2093**	**0.2144**	**-0.0532**	**-0.0554**	**-0.0181**	*0.0044*	*0.0004*	*0.0040*	**-0.0128**
Reviews: Creation Month	**0.2094**	**0.2144**	**-0.0531**	**-0.0554**	**-0.0181**	*0.0044*	*0.0005*	*0.0042*	**-0.0130**
Reviews: Creation Year	**0.1939**	**0.1989**	**-0.0509**	**-0.0477**	**-0.0154**	*0.0041*	*0.0012*	*0.0050*	**-0.0109**

Table 4.6: Results of Feature Correlation Tests on Training Set, Word-Class Features (Vitamins)

	Date Variance	Avg. Word Length	Avg. Sentence Length	Avg. Review Length	Word Perplexity	POS Perplexity
Products: Star-Rating Skewedness (5- and 1-star)	**-0.1242**	**0.0933**	**-0.0709**	*-0.0187*	*0.0217*	**-0.1784**
Products: Star-Rating Skewedness (5-star)	**-0.1468**	**0.1093**	**-0.0808**	*-0.0390*	*0.0255*	**-0.1807**
Products: Review Creation Date Variance	N/A	**-0.0884**	**0.1971**	**0.1542**	*-0.0064*	**0.2979**
Reviews: Creation Date	N/A	**0.0880**	**-0.2048**	**-0.1759**	N/A	N/A
Reviews: Creation Month	N/A	**0.0876**	**-0.2046**	**-0.1759**	N/A	N/A
Reviews: Creation Year	N/A	**0.0839**	**-0.1908**	**-0.1623**	N/A	N/A

Table 4.7: Results of Feature Correlation Tests on Training Set, Non-Lexical Features (Men's Underwear)

	Sentiment	Positive Sentiment	Negative Sentiment	Pronouns	"Scientific" Words	Imperatives	Comparatives	Null Comparatives	Brand Names
Products: Star-Rating Skewedness (5- and 1-star)	**0.1315**	**0.1346**	**-0.0600**	**-0.0715**	*0.0341*	*-0.0024*	**-0.0915**	**-0.0787**	*-0.0374*
Products: Star-Rating Skewedness (5-star)	**0.1433**	**0.1476**	**-0.0744**	**-0.1022**	*0.0288*	*0.0055*	**-0.0760**	**-0.0610**	*-0.0404*
Products: Review Creation Date Variance	**-0.1618**	**-0.1708**	**0.1280**	0.0576	*0.0354*	*-0.0005*	*0.0294*	*0.0276*	**0.1485**
Reviews: Creation Date	**0.2575**	**0.2619**	**-0.0607**	**-0.0917**	-0.0098	**-0.0170**	**-0.0228**	**-0.0212**	**-0.2606**
Reviews: Creation Month	**0.2575**	**0.2619**	**-0.0607**	**-0.0917**	-0.0098	**-0.0170**	**-0.0227**	**-0.0211**	**-0.2605**
Reviews: Creation Year	**0.2368**	**0.2403**	**-0.0530**	**-0.0823**	*-0.0082*	**-0.0171**	**-0.0210**	**-0.0188**	**-0.2482**

Table 4.8: Results of Feature Correlation Tests on Training Set, Word-Class Features (Men's Underwear)

thresholds to generate the gold-standard class labels. The approach and justification for these experiments is described in more depth in section 3.3.

We used the training data for all product categories to determine optimal model parameters and explore the effectiveness of various features. Then, we trained classifiers on the training data for each product category, as well as on the full training dataset. We tested each classifier on the corresponding test set for its category. We also tested each classifier on the full test set, to determine whether feature correlations within a given product category would hold across categories.

All classification experiments were performed with MALLET (McCallum, 2002) using a maximum entropy classifier. All experiments on the training set used 4-fold cross-validation to split the data into training and validation folds.

4.2.1 Training & Tuning the Models

We explored using both the percentage of 5- and 1-star reviews and the percentage of 5-star reviews as a labeling heuristic, as well as varying the minimum and maximum thresholds. We ultimately determined that using the percentage of 5-star reviews yielded a higher classification accuray. To put it another way, we found the 5-star reviews of products with a high percentage of 5-star reviews to be more unique than the 5-star reviews of products with a marked bimodal distribution of review star ratings, at least in terms of the features that we used for classification. Thus, the labeling for all experiments in subsequent sections is based on the percentage of the product's reviews that are 5-star. Similarly, we settled on minimum and maximum thresholds based on trial and error. This process involved a tradeoff between a blurring of the boundary between classes (as we set the thresholds closer together) and the diminishing size of the dataset (as we set them farther apart).

Parameter Tuning

We performed a series of experiments on the full dataset using word unigrams as features to determine the optimal parameters for subsequent experiments.

Max. Number of Reviews	Instances per Class	Min. Threshold for Suspicious Class	Max. Threshold for Trustworthy Class	Avg. Training Accuracy	Avg. Test Accuracy
25	2821	0.8571	0.5696	**67.56%**	**66.71%**
50	2821	0.8571	0.5696	**67.59%**	**66.89%**
100	2821	0.8571	0.5696	**67.74%**	**66.55%**
150	2821	0.8571	0.5696	**67.56%**	**66.75%**
200	2821	0.8571	0.5696	**67.95%**	**66.50%**

Table 4.9: Effects of Maximum Number of Reviews per Product on Classification Accuracy (Word Unigram Features, All Products)

To begin, we experimented with sorting each product's reviews by length and keeping only the top N longest reviews. Results are shown in table 4.9. We determined that using the 50 longest reviews was sufficient; adding additional reviews did not substantially improve classification accuracy. We used this parameter setting for all subsequent experiments.

In addition, we tried discarding reviews below a minimum length. Very short reviews could mute the signal of certain linguistic features contained in longer reviews. Through a series of experiments described in table 4.10, we determined that a minimum review length of 10 was optimal.

We also experimented with feature pruning in the experiments using word unigram features, for which the feature space is very large. We determined through the experiments described in table 4.11 that pruning unigram features below a certain value did not help classification accuracy, and could even hurt it. Thus, we set this parameter to 0 for subsequent experiments. However, we determined that using a stopword list when generating

Min. Review Length	Instances per Class	Min. Threshold for Suspicious Class	Max. Threshold for Trustworthy Class	Avg. Training Accuracy	Avg. Test Accuracy
1	2821	0.8571	0.5696	**68.78%**	**66.80%**
10	2821	0.8571	0.5696	**68.73%**	**67.35%**
20	2821	0.8571	0.5696	**67.59%**	**66.89%**
40	2821	0.8571	0.5696	**62.29%**	**61.56%**
60	2821	0.8571	0.5696	**58.91%**	**58.56%**
80	2821	0.8571	0.5696	**59.09%**	**59.09%**
100	2821	0.8571	0.5696	**60.03%**	**59.96%**

Table 4.10: Effects of Minimum Review Length on Classification Accuracy (Word Unigram Features, All Products)

Min. Value for Unigram Features	Instances per Class	Min. Threshold for Suspicious Class	Max. Threshold for Trustworthy Class	Avg. Training Accuracy	Avg. Test Accuracy
0	2821	0.8571	0.5696	**67.59%**	**66.89%**
0.0005	2821	0.8571	0.5696	**67.64%**	**66.87%**
0.0001	2821	0.8571	0.5696	**67.68%**	**66.90%**
0.00005	2821	0.8571	0.5696	**67.76%**	**66.24%**
0.00001	2821	0.8571	0.5696	**67.61%**	**66.27%**

Table 4.11: Effects of Feature Pruning on Classification Accuracy (Word Unigram Features, All Products)

word unigram features had a positive impact on classification accuracy; see table 4.12.

Amazon.com reviews have a "Verified Purchase" purchase field that reviewers can choose to activate if Amazon.com is able to verify that the product was purchased on the site. We collected the value for this field during screen scraping, but subsequently determined that filtering for reviews labeled "Verified Purchase" had a negative effect on classification accuracy; see table 4.13. This may be because many deceptive reviewers did not actually purchase the product they are reviewing.

We used the parameters calculated above for the final experiments on all datasets.

Feature Selection

To determine the optimal combination of features, we performed classification experiments on the full dataset using different combinations of features. Results are given in tables 4.14-4.15. All experiments used 5-star rating percentage thresholds of 0.8571 and 0.5696, and

	Instances per Class	Min. Threshold for Suspicious Class	Max. Threshold for Trustworthy Class	Avg. Training Accuracy	Avg. Test Accuracy
Keep Stopwords	2821	0.8571	0.5696	**67.59%**	**66.89%**
Remove Stopwords	2821	0.8571	0.5696	**71.79%**	**70.19%**

Table 4.12: Effects of Stopword Removal on Classification Accuracy (Word Unigram Features, All Products)

	Instances per Class	Min. Threshold for Suspicious Class	Max. Threshold for Trustworthy Class	Avg. Training Accuracy	Avg. Test Accuracy
Filter for Verified Purchase	2821	0.8571	0.5696	**55.61%**	**55.55%**
Do Not Filter	2821	0.8571	0.5696	**67.59%**	**66.89%**

Table 4.13: Effects of Filtering for Verified Purchase on Classification Accuracy (Word Unigram Features, All Products)

2821 products per class.

In addition to word and POS n-gram features, we used features that exhibited a high correlation score in the feature correlation tests. To better ensure that we were capturing textual elements that would transfer across product categories, we only used features that we felt confident we could explain. For example, we excluded the `comparatives` feature, which consistently exhibits a significant correlation with the star-rating skewedness, but the opposite correlation from that which we would expect based on the advertising literature.

We distinguish between experiments that use unigram features and those that do not. The reason for this is that unigrams could encode information about the type of product being reviewed. If the star rating distribution for a particular type of product tends to be more skewed than other types, a classifier could learn this relationship, and predict the "Suspicious" class for this product based on unigram features. However, developing a model that is able to classify products as suspicious or trustworthy with high accuracy based solely on the type of product does not answer our research question as to whether reviews for deceptive products exhibit different stylistic elements than reviews for trustworthy products.

4.2.2 Results on Held-Out Test Set

We trained classifiers on the full training set for each category of products, as well as the combined training set containing all categories of products. We used the optimal parameter settings discovered above. Based on the results above, we determined a combination of POS unigram and bigram features to be the optimal feature combination excluding word n-gram features. We believe that word n-gram features did not reflect actual stylistic differences, since they seemed to be capturing types of products (see table 5.1). Furthermore, we felt that lexical features would not transfer well to new categories of products.

Results on Each Product Category

We tested each classifier on the corresponding test set for its category of products. Results are given in table 4.16. The minimum and maximum thresholds correspond with the mini-

Word Unigram Features	Word Bigram Features	Function Word Features	Sentiment Word Features	POS Unigram Features	POS Bigram Features	Avg. Word Length Feature	Avg. Review Length Features	POS Perplexity Feature	Avg. Training Accuracy	Avg. Test Accuracy
X									**67.59%**	**66.89%**
X	X								**69.84%**	**68.15%**
X	X	X	X	X	X	X	X	X	**60.95%**	**60.76%**
X	X	X	X	X	X	X	X		**64.41%**	**63.08%**
X	X	X	X	X	X	X		X	**60.78%**	**60.40%**
X	X	X	X	X	X		X	X	**63.05%**	**62.02%**
X	X	X	X	X		X	X	X	**56.67%**	**56.64%**
X	X	X	X		X	X	X	X	**57.86%**	**57.55%**
X	X	X		X	X	X	X	X	**55.15%**	**55.09%**
X	X		X	X	X	X	X	X	**57.76%**	**57.43%**
X		X	X	X	X	X	X	X	**55.09%**	**55.19%**
	X	X	X	X	X	X	X	X	**53.93%**	**53.93%**

Table 4.14: Results of Classification Experiments Using Different Feature Sets on Training/Validation Set (All Products, Include Word N-grams)

Function Word Features	Sentiment Word Features	POS Unigram Features	POS Bigram Features	Avg. Word Length Feature	Avg. Review Length Features	POS Perplexity Feature	Avg. Training Accuracy	Avg. Test Accuracy
X							**60.95%**	**60.76%**
	X						**64.49%**	**63.08%**
		X					**60.78%**	**60.40%**
		X	X				**63.05%**	**62.02%**
		X				X	**56.67%**	**56.65%**
		X	X			X	**57.86%**	**57.55%**
X		X	X			X	**57.76%**	**57.43%**
X		X	X	X	X	X	**55.09%**	**55.19%**
X	X	X	X	X	X	X	**53.93%**	**53.93%**

Table 4.15: Results of Classification Experiments Using Different Feature Sets on Training/Validation Set (All Products, No Word N-grams)

mum percentage of reviews that were 5-star for products in the "Suspicious" class and the maximum percentage of reviews that were 5-star for products in the "Trustworthy" class, respectively.

Transferability of Models

To determine the transferability of models developed on a particular category of product, we tested each classifier trained on a single category of products on the test set for all products. Results are given in table 4.17.

4.3 Human Performance Experiments

An overview of Turkers' responses to the question about review trustworthiness is given in tables 4.18 and 4.19.

We calculated inter-rater agreement using Fleiss' kappa. Inter-rater agreement at the review level and the sentence level is given in table 4.20.

In these HITs, we asked Turkers to rate reviews as "Very Trustworthy," "Somewhat Trustworthy," "Somewhat Suspicious," or "Very Suspicious" in order to gain additional insight into their perception of the reviews. However, because we were ultimately exploring binary classification and responses of the same polarity should be treated as more similar than responses of opposite polarity, we treated a response of "Very Trustworthy" or "Somewhat Trustworthy" (or "Very Suspicious" or "Somewhat Suspicious") as equivalent in our calculation of inter-rater agreement.

In addition to inter-rater agreement for the review-level labels, we calculated inter-rater agreement for the sentence-level labels. For each sentence, Turkers could have selected it as contributing to their perception of the review as trustworthy, contributing to their perception of the review as suspicious, or neither. Since Turkers were not required to make a selection for each sentence but rather simply to select sentences that influenced their rating of the review, many sentences were not selected by any Turkers and thus were assigned the same

Product Category	Instances per Class, Training	Min. Threshold, Training	Max. Threshold, Training	Instances per Class, Test	Min. Threshold, Test	Max. Threshold, Test	Training Accuracy	Test Accuracy
All	2821	0.8571	0.5696	706	0.8535	0.5625	**63.84%**	**63.03%**
Bath & Body	230	0.8542	0.54	60	0.852	0.54	**70.43%**	**64.16%**
Cleaning	143	0.849	0.49	60	0.8168	0.5	**69.58%**	**52.50%**
Cough	713	0.92	0.79	198	0.925	0.843	**62.97%**	**58.58%**
Fragrance	255	0.8405	0.651	66	0.838	0.65	**60.78%**	**59.84%**
Hair Products	143	0.75	0.55	47	0.75	0.612	**66.43%**	**59.57%**
Herbal Supplements	462	0.849	0.65	111	0.84	0.615	**59.84%**	**55.86%**
Women's Lingerie	90	0.7	0.427	24	0.7	0.45	**69.44%**	**70.83%**
Skin Products	140	0.8	0.575	30	0.8	0.579	**67.14%**	**68.33%**
Men's Underwear	645	0.7272	0.509	167	0.7	0.5	**58.22%**	**53.89%**
Vitamins	469	0.8572	0.63	118	0.8572	0.625	**62.90%**	**59.75%**

Table 4.16: Results of Category-Specific Classification Experiments (POS Unigram and Bigram Features)

Product Category, Training	Product Category, Test	Instances per Class, Training	Min. Threshold, Training	Max. Threshold, Training	Instances per Class, Test	Min. Threshold, Test	Max. Threshold, Test	Training Accuracy (Category-Specific)	Test Accuracy (Full Test Set)
Bath & Body	**All**	230	0.8542	0.54	706	0.8535	0.5625	**70.43%**	**58.50%**
Cleaning	**All**	143	0.849	0.49	706	0.8535	0.5625	**69.58%**	**57.51%**
Cough	**All**	713	0.92	0.79	706	0.8535	0.5625	**62.97%**	**55.10%**
Fragrance	**All**	255	0.8405	0.651	706	0.8535	0.5625	**60.78%**	**53.90%**
Hair Products	**All**	143	0.75	0.55	706	0.8535	0.5625	**66.43%**	**53.61%**
Herbal Supplements	**All**	462	0.849	0.65	706	0.8535	0.5625	**59.84%**	**57.15%**
Women's Lingerie	**All**	90	0.7	0.427	706	0.8535	0.5625	**69.44%**	**53.47%**
Skin Products	**All**	140	0.8	0.575	706	0.8535	0.5625	**67.14%**	**54.18%**
Men's Underwear	**All**	645	0.7272	0.509	706	0.8535	0.5625	**58.22%**	**54.53%**
Vitamins	**All**	469	0.8572	0.63	706	0.8535	0.5625	**62.90%**	**57.44%**

Table 4.17: Performance of Category-Specific Classifiers on Full Dataset (POS Unigram and Bigram Features)

Label	Number of Responses
"Very Suspicious"	30
"Somewhat Suspicious"	31
"Somewhat Trustworthy"	39
"Very Trustworthy"	44

Table 4.18: Perception of Review Trustworthiness: Distribution of Responses by HIT

Majority Label	Number of Reviews
"Very Suspicious" or "Somewhat Suspicious"	21
"Very Trustworthy" or "Somewhat Trustworthy"	27

Table 4.19: Perception of Review Trustworthiness: Distribution of Majority Labels by Review

null label, so it is possible that the inter-annotator agreement at the sentence level may be higher than it would be if we had required a selection for each sentence.

Table 4.21 gives a summary of common reasons Turkers gave for rating a review as suspicious or trustworthy, respectively.

4.4 Discourse-Based Review Classification

4.4.1 Sentence-Level Classification

Sentence-tagging experiments were conducted over a dataset of 441 labeled sentences (4 classes), using word and POS n-gram features. Results are given in table 4.22.

Label	Fleiss' Kappa
Review-Level	0.431
Sentence-Level	0.340

Table 4.20: Inter-Rater Agreement for Human Performance Experiments

Label	Explanations
Suspicious	1. Mention of the product or brand name 2. Copy perceived as "too positive" or "over-praising" 3. Too many exclamation points or capital letters 4. Too much knowledge about the product (e.g., listing/describing ingredients, listing uses for the product, extensive detail about product usage regimen) 5. No description at all of the reviewer's own experience using the product 6. Commands (e.g., "Do yourself a favor and buy this!!!!!!") 7. Catchy phrases (e.g., containing parallelism, as in "so, simple, yet so effective") 8. Pseudo-scientific marketing phrases (e.g., "healthy and vibrant," "organic," "astringent," "residue," "super slick oil")
Trustworthy	1. Any mention of a negative aspect of the product 2. Mention of competing products 3. Misspellings, grammar errors, "disorganized," "casual tone" 4. Emoticons 5. Shorter review length 6. Personal details or account of reviewers' experience using the product

Table 4.21: Common Explanations for Human Perceptions of Review Trustworthiness

Average Training Accuracy	Average Test Accuracy
75.7%	47.6%

Table 4.22: Results of Sentence Tagging Experiments

Model	Average Training Accuracy	Average Test Accuracy
Baseline model (word/POS unigrams/bigrams)	96.1%	43.8%
Baseline model + sentence-tag unigrams	67.6%	62.5%
Baseline model + sentence-tag bigrams	83.3%	55.4%
Baseline model + sentence-tag unigrams/bigrams	71.4%	59.2%
Just tag unigrams/bigrams	73.0%	59.6%

Table 4.23: Results of Review Classification Experiments

4.4.2 Review-Level Classification

A small set of experiments were conducted using sentence-tag n-gram features to classifiy individual reviews as suspicious or trustworthy according to human judgments (in contrast with the attempts we made to detect actual deception in sections 3.2-3.3). Because of the poor performance of the current sentence classifier, we used manually generated sentence labels for these experiments.

Experiments were conducted using 4-fold cross-validation. The average training and test accuracies for each experimental configuration are reported in table 4.23.

Chapter 5
DISCUSSION

While our experiments yielded mixed results, several interesting trends emerged. These observations, as well as our observations about what did not work, could inform future research in the characterization and detection of deceptive reviews.

5.1 Feature Correlation Tests

Our exploration of feature correlation revealed a number of interesting trends (see tables 4.1-4.8). It is important to note that while none of these correlation values appear particularly high, many of them are statistically significant given that they were calculated across hundreds or thousands of data points (in the case of the product and review experiments, respectively).

We observed a consistent positive correlation between review creation date and the density of positive sentiment words in 5-star reviews. Based on these calculations, we can say that 5-star reviews appear to be growing *more* positive with time. Given the assumption that the volume of fake reviews is most likely increasing over time, this correlation is consistent with our hypothesis that fake 5-star reviews tend to be more positive than genuine 5-star reviews.

Furthermore, we observed a positive correlation between the skewedness of a product's star-rating distribution (namely, a high percentage of 5-star reviews, or a high percentage of 5- and 1-star reviews) and the density of positive sentiment words in 5-star reviews. This is a very interesting trend; 5-star reviewers tend to write more positively about products that tend to elicit stronger opinions in reviewers on average. This is not only the case for "popular" products (those that tend to have a high percentage of 5-star reviews); the

correlation also manifests (albeit less strongly) for "polarizing" products (those that tend to have a high percentage of 5- and 1-star reviews). These observations highlighted positive sentiment features as a potentially useful feature for the product classification experiments described in sections 3.3 and 4.2.

We also observed a significant negative correlation between average sentence length and star-rating skewedness and between average review length and star-rating skewedness. This is consistent with our observation that suspicious products often have a large number of extremely short 5-star reviews. On the other hand, there was a significant positive correlation between average word length and the percentage of 5-star (or 5- and 1-star) reviews. We hypothesize that this is because, despite the previously mentioned trend of suspicious products having many very short reviews (which also tend to contain very short words), the very longest, most detailed reviews with the longest words also tended to show up in suspicious products. Since word length was normalized by the total number of words across reviews without regard for review boundaries, the average word length in these very long reviews would outweigh the average word length in the extremely short reviews. We observed the same correlations for reviews on the time axis: Reviews and sentences appear have become shorter on average over time, while the average word length has increased.

There was a significant negative correlation between the perplexity of a POS-unigram language model and star-rating skewedness across the four datasets. This is consistent with our hypothesis that suspicious products tend to exhibit less language variaton across reviews, as many of the reviews may be written by the same author. (The correlation between word perplexity and star-rating skewedness was not significant.) There was also a significant positive correlation between the POS perplexity and date variance across all four datasets, meaning that products with reviews created over a longer period of time tend to have more varied language in their reviews. While this is consistent with our hypotheses about perplexity and date variance correlating positively with review trustworthiness, it is also possible that a shift in the way reviewers write over time contributes to this trend.

There was a negative correlation between star-rating skewedness and review creation date

variance in the full dataset, and in general, the correlations between review creation date variance and the various features explored were the reverse sign of the correlations between review creation date or star-rating skewedness and the other features explored. This is consistent with our hypothesis that products with a highly skewed distribution of review star ratings will also have a greater number of reviews produced in a short amount of time (i.e., a lesser degree of variance in the review creation date). We believe that both of these features are indicators of suspicious products.

As discussed in section 3.1, the two clothing categories (women's lingerie and men's underwear) were included as a control, based on the hypothesis that there would be less suspicious activity in these categories than in cosmetics and health supplements. However, the correlations in the "Men's Underwear" dataset resemble the correlations in the other three datasets in terms of significance, sign, and strength. Further research is necessary to determine whether this is indicative of suspicious activity existing even in product categories where one would not suspect it, or if the trends observed here are not actually correlated solely with suspicious activity, but with other causes as well.

Unfortunately, our hand-crafted advertising-based features (`imperatives`, `comparatives`, `null_comparatives`, and `name_mentions`) did not show promise for the identification of suspicious products and reviews. There are a number of possible reasons for this. For one thing, `comparatives` and `null_comparatives` relied on a POS-tagger. A random sample of the POS-tagger's output contained a number of errors, which may be in part due to the fact that it was trained on text from a different domain. In addition, we relied on heuristics for identifying imperatives and null comparatives, such as searching for the base form of the 100 most common verbs at the start of a sentence (in the case of imperatives) and searching in a window after a comparative adjective for the word *than* (in the case of null comparatives). Our `name_mentions` feature was only activated by exact string matches with the product, brand, and seller name; it was not able to capture misspellings. These methods are not comprehensive, and it is possible that the features were not triggered in a number of cases when they should have been. However, the extremely weak correlation values (and the hand-

ful of correlation values with the opposite sign of what we would expect) do suggest that features that give a more comprehensive picture of a product's reviews, such as sentiment word density or POS-tag distributions, are more likely to be useful for identifying products with a skewed review star-rating distribution than these hand-crafted features designed to capture specific cases.

Neither the density of personal pronouns nor the density of scientific words exhibited significant correlations in the directions we expected. In fact, these features exhibited significant negative correlations with review star-rating skewedness in many cases. This suggests that fake reviews do not necessarily exhibit the content elements that arouse suspicion in humans.

5.2 Text-Based Product Classification

In discussing the results of the product classification experiments, it is important to keep in mind that our labels are based on the assumption that products containing deceptive reviews tend to have a skewed star-rating distribution. While previous research, such as that of Feng et al. (2012b), leads us to believe that this is a reasonable assumption, we do not have any way of verifying that it is in fact true. Therefore, a particular classifier's high performance means it was effective in detecting products with skewed star-rating distributions. Similarly, if a feature worked well (or poorly) for classification, the most we can say is that the feature was effective (or not effective) for identifying products with skewed star-rating distributions. To make the statement that these classifiers or features are in fact effective in detecting fake reviews, we would need further evidence that skewed star-rating distributions and deceptive products are in fact correlated. Unfortunately, it is extremely difficult to identify fake reviews (or products with suspicious activity) with any degree of certainty.

5.2.1 Parameter Tuning

Finding the optimal parameter settings for these experiments involved facing the tradeoff between sufficiently distinct classes and sufficient data. As the minimum and maximum 5-

star review percentage thresholds for the two classes were moved farther apart, the differences between the two classes became more pronounced, but the dataset became smaller. This could have contributed to classification working less well for certain categories of products.

Furthermore, in order for classification to work well, the product category in question needs to contain a mixture of products for which many reviews are deceptive and products without a large number of deceptive reviews. Lower classification accuracy for a particular category does not necessarily mean that there are fewer fake reviews in this category, but rather that the clusters of products with high and low percentages of 5-star reviews are not distinct. For example, classification accuracy on the test set was substantially lower for Herbal Supplements, at 55.86%, than for the full dataset, at 63.03%. We suspect that this is actually because a large number of products in this category contain fake reviews, even those products that do not have a particularly skewed star rating distribution. This means that the contrast between the classes is not particularly extreme, and it is difficult for the classifier to distinguish between the two classes. We hypothesize that the categories with the best performance (e.g., Bath & Body; Skin Products) are categories that contain a mixture of products that exhibit suspicious activity and products that do not.

5.2.2 Feature Selection

We performed experiments using various combinations of feature sets on the training/validation set for all products to determine the optimum combination of feature sets (see tables 4.14-4.15). We achieved the highest classification accuracy with word unigram and bigram features, followed by just word unigram features. These models outperformed other models by a significant margin. Adding feature sets to these models hurt classification accuracy in all cases. However, in examining the model's most highly weighted features, given in table 5.1, we determined that the classifier was utilizing information about product type. This suggests that certain types of products may be more likely to have extremely skewed star ratings and, by extension, possibly more suspicious activity. However, since we are most interested in detecting stylistic differences in the reviews of suspicious products, we omitted

these features for subsequent experiments, despite their high performance.

After omitting the full set of word unigram and bigram features, we achieved optimal classification accuracy (63.08%) with sentiment word features. This suggests that the 5-star reviews of products with extremely skewed star rating distributions are more positive on average than the 5-star reviews of products with more uniform distributions. This is an interesting finding, and suggests that star-rating distributions may in fact be correlated with product suspiciousness. At the very least, we are able to make the statement that 5-star reviewers write *more* positively about products with a high percentage of 5-star reviews than about products with a more uniform star rating distribution.

We also achieved classification accuracies greater than 60% with a combination of POS unigram and bigram features (62.02%), with function word features (60.76%), and with just POS unigram features (60.40%). All of these features can be said to characterize writing style, and the fact that we were able to achieve classification accuracy more than 20% better than a random baseline suggests that reviewers of products with highly skewed star rating distributions share a distinctive writing style. Adding feature sets to each of these models hurt classification accuracy. In many cases, it appears that the classifier assigned too much weight to one-off features such as POS perplexity and average-length features, distorting the more reliable signal provided by POS or lexical feature sets.

5.2.3 Results on Held-Out Test Set

In our analysis of highly weighted features for individual models, we determined that the classifier appeared to be benefiting from information about the category or subcategory of product encoded in lexical features. While the classifiers trained only on function and/or sentiment word features appeared more category-agnostic, words related to product types do appear in these lexicons and can be assigned very high weights, as can be seen for sentiment-word features in table 5.1. Thus, our decision to omit all lexical features from our final experiments had two motivations: 1) the concern that lexical features would not transfer well across product categories, and 2) our interest in exploring whether deceptive reviews

Feature Sets	Top Features for Suspicious Class	Top Features for Trustworthy Class
Word Unigrams	this, vitamin, cold, cough, take, product, throat, taking, me, best, been, taste, years, works, d	the, but, i, fit, hair, t, they, comfortable, size, bra, very, and, are, not, look
Word Unigrams & Bigrams	this, product, cold, take, cough, vitamin, for, taking, throat, taste, best, me, START-this, we, years	the, i, hair, fit, but, comfortable, t, they, size, bra, and, up, that, perfect, not
Function Words	this, for, me, we, been, from, day, at, using, always, my, our, by, of, have	the, i, they, but, and, just, up, are, not, very, these, would, more, that, little
Sentiment Words	cold, works, sore, best, sick, recommended, favorite, smell, allergies, pain, symptoms, effective, love, wonderful, relief	comfortable, perfect, like, well, nice, perfectly, great, clean, right, soft, fine, enough, support, top, warm
POS Unigrams	NNP, NNS, PRP$, JJS, VBZ, VBN, NNPS, IN, -NONE-, PDT, NN, LS, WP$, UH, TO	RB, JJ, CC, DT, VBD, VBP, VB, PRP, RP, CD, MD, JJR, RBR, WP, WDT
POS Unigrams & Bigrams	NNP, NNS, IN-NNP, NN-STOP, PRP$, NNS-STOP, PRP-IN, IN-NN, JJS, VBZ-DT, DT-VBZ, VBZ, JJ-NNS, VBG, START-NNP	RB, VBD, CC, PRP-VBD, JJ, IN-DT, CD, RB-JJ, DT-NN, DT-NNS, JJ-STOP, RB-RB, DT, RP, JJ-CC

Table 5.1: Highly Weighted Features for Various Feature Set Combinations (All Products)

exhibited stylistic differences independent from product type.

In the case of all categories, a model trained on POS unigrams and bigrams was able to predict the product label with a higher accuracy than a baseline model using random guessing (see table 4.16). The accuracy of a model trained on the training dataset for all products and tested on the full test set was 63.03%. However, the accuracy was substantially better for some product categories than others. The datasets for bath & body products, women's lingerie, and skin products had better accuracies than the model trained and tested on the full dataset, while the models trained and tested on cleaning products and men's underwear had the lowest accuracies, at less than 55%. We suspect that this is because cleaning products and men's underwear have a relatively low concentration of products with deceptive reviews, and therefore, the difference between products with a skewed star rating distribution and those with a more uniform distribution was relatively unpronounced. In contrast, we believe that the datasets corresponding with bath & body products, women's lingerie, and skin products contained a more uniform mixture of deceptive and trustworthy products.

We hypothesized that herbal supplements would be a category for which the concentration of fake reviews would be quite high, and the classification accuray for this category is relatively low, at 55.86%. We believe that this model suffers from a similar problem as the models trained on cleaning products and men's underwear, only in this case, the issue is not a dearth of deceptive products in the dataset, but rather, a dearth of trustworthy products.

The most highly weighted POS n-gram features for each category, for each class, are given in table 5.2. In future work, we would like to explore whether these features have similar predictive power for detecting deception and/or advertising intent in other domains.

5.2.4 Transferability of Models

While none of the models trained on specific categories of products failed to beat a random baseline when tested on the full dataset, the classification accuracy on the full test set was lower than the classification accuracy on the test set for the category on which the model

Category	Top Features for Suspicious Class	Top Features for Trustworthy Class
All	NNP, NN-STOP, IN-NNP, NNS, NNS-STOP, PRP-IN, PRP$, START-NNP, VBG, JJS	RB, CC, PRP-VBD, CD, VBD, NNP-NNP, JJ, IN-DT, VB, DT-NN
Bath & Body	NNP-NN, NNS, NN-STOP, NN-VBZ, START-NNP, NNS-STOP, IN-NNP, VBZ, START-NNS, PRP-IN	NNP-NNP, PRP, VB, TO-VB, TO, VBD, CC-PRP, PRP-VBD, CD, PRP$-NN
Cleaning	NNP, NNS, NN-NN, START-NNP, PRP$, NN-NNS, NNS-STOP, JJS, NN-STOP, NNP-VBZ	DT, DT-NN, IN, IN-DT, PRP, VBD, RB, CD, TO, JJ
Cough	START-NNP, NN-STOP, START-DT, NNP, VBP-DT, NNP-VBD, START-NN, PRP-IN, JJ-NN, PRP-STOP	IN, RB, CD, VB, PRP, VBD, CC, VBZ, TO, PRP-VBD
Fragrance	PRP, NN-STOP, VBN, RB, START-NNP, PRP-IN, WP, VBP-VBN, PRP-VBZ, VBP	DT, JJ, NN, IN-DT, IN, DT-JJ, CC, DT-NN, JJ-NN, VB
Hair Products	NN, VBZ, NNP, CC, NN-STOP, PRP$, PRP$-NN, PRP-VBZ, VBZ-DT, START-NNP	VBD, PRP-VBD, IN, NNS, VB, CD, TO, TO-VB, IN-DT, DT-NNS
Herbal Supplements	NNP, START-NNP, NN, NNS, NNS-STOP, NN-STOP, NN-NN, JJ, IN-NNS, NNP-NN	RB, IN, CD, VB, PRP-VBD, TO-VB, IN-PRP, NN-PRP, RB-RB, VBG
Women's Lingerie	VBP, PRP-VBP, JJ, NNS, NNP, RB, VBP-RB, START-DT, NN-CC, VBP-VBN	VBD, NN, NN-NN, PRP-VBD, PRP, PRP-VBZ, PRP$, CC-PRP, VBD-DT, PRP$-NN
Skin Products	NN, NNP, NN-STOP, START-NNP, VBZ, START-PRP, NN-NN, START-DT, NN-CC, PRP$	VB, TO, PRP, VBD, TO-VB, PRP-VBP, NNS, IN-PRP, VBP, DT-NNS
Men's Underwear	NN-CC, NNP, CC, VBG, CC-DT, RB-IN, START-NNP, VBP-VBG, START-NNS, JJ-NN	PRP-VBZ, VBN, VBZ, PRP, NN-NNP, DT-NN, DT, VBZ-DT, VBD, IN
Vitamins	NNP, START-DT, JJ, START-NNP, JJ-TO, PRP$, VBN, VBZ, IN-NNP, DT	PRP, RB, CC, VB, NN-NN, PRP-VBD, NN, NN-CC, VBD, IN

Table 5.2: Highly Weighted Features for Various Product Categories

was trained 7 out of 10 times, and lower than the classification accuracy of the model trained on the full training set 10 out of 10 times. This suggests that while there are cross-category trends in the way that 5-star reviewers write about products with highly skewed star ratings (as can be seen from the test accuracy achieved with the model trained on the full dataset), there are stylistic differences in the way that reviewers write about different products, and classifiers trained on specific categories of products may not perform as well when tested on products from other categories.

5.3 Human Performance Experiments

Two interesting observations emerged from our MTurk experiments. The first is that human annotators have a fairly high level of agreement regarding suspiciousness or trustworthiness at the review level. While it is possible that the reviews we picked for these experiments were particularly extreme, we observed that the judgments of Turkers frequently did not align with our own. On average, Turkers tended to be more trusting than we were; the majority label was "Trustworthy" for 27 out of 48 reviews, and 83 out of 144 responses indicated that the review in question was trustworthy (in a dataset that we perceived to contain only 50% trustworthy reviews).

Our second observation based on these experiments is that human judgments of fake reviews do not align with the predictions made by our models. There was not a significant correlation between reviews' trustworthiness labels and any of the features in the correlation tests described in sections 3.2, 4.1, and 5.1. Past research suggests that humans are not good at identifying fake reviews (Ott et al., 2011). In the case of our experiments, we cannot say for certain that the fact that our model's predictions do not align with human's predictions means that our model is outperforming humans, given that our gold standard is based on a heuristic. Nevertheless, it is interesting to examine the ways in which human predictions and the predictions made by our model differ. While humans appear to be less sensitive to the shallow textual cues around which we built our model, they are able to draw conclusions based on compex reasoning that our model cannot. Consider the following review from the

Herbal Supplements dataset:

> *I'll start by saying that I'm 32 years old, in good health, eat well and I exercise. I also have 5 children under the age of 9 that I homeschool. I've suffered from depression/anxiety/stress and exhaustion. I've tried different herbs over the years, meds never mixed well with my system and I try to do things naturally when possible...*
>
> *When my bottle of L-theanine ran out, I began searching for something new to try. This looked promising, but I'm addicted to product reviews and couldn't find any online anywhere. It was inexpensive enough for me to try, so I gave it a shot. Today is Wednesday, and I started taking them last Friday. I take 2 for breakfast, 1 for lunch. These are working incredibly well for me. Like, miraculously well. I wish I could explain just how exhausted I've been the last 2 years. Our house has been a disaster. It's been at that point where it was so overwhelming not knowing where to start, that nothing gets done. I could barely get out of bed some mornings, and I think my mood was partially caused by my lack of energy.*

Two out of three human annotators commented on the incongruency between the reviewer's initial assertion that he/she "eat(s) well and... exercise(s)" and the subsquent statement, "I wish I could explain just how exhausted I've been the last 2 years... I could barely get out of bed some mornings, and I think my mood was partially caused by my lack of energy." This is the sort of reasoning that is very difficult for a computer to do. On the other hand, it is something that humans can do somewhat reliably. This example highlights the point that while automated methods for detecting online opinion fraud can pick up on textual signals that humans do not reliably detect, humans are able to identify contradictions in a storyline that are much more difficult for a computer to detect.

5.4 Discourse-Based Review Classification

It is difficult to interpret the results of the discourse-based review classification experiments because of the limited nature of the dataset. This was mostly an exercise in experimental design, as hand-labeling reviews and sentences is very expensive.

5.4.1 Sentence-Level Classification

While the performance of the sentence classifier was not good enough to use the automatically generated labels for review classification, it was substantially better than a naive baseline that always chooses the majority class, and therefore we believe that automatic sentence-tagging shows promise. In future experiments, we hope to improve sentence tagging and review classification through significantly augmenting the dataset. We also believe that we could potentially improve sentence tagging through adding rules.

5.4.2 Review-Level Classification

In the review classification experiments performed, using sentence tag n-grams improves the performance of the classifier slightly, though the baseline classifier has essentially the same performance as a random baseline and the signal is not strong even after adding tag n-grams. We suspect that there are two main issues behind the poor results: an insufficient amount of data, and the inability of the current sentence tags to adequately capture content that arouses suspicion or trust. In order to assess this approach more confidently, we would need to conduct experiments on a much larger scale, employing some combination of crowdsourcing and/or automatic labeling of sentences using heuristics to build a dataset. A larger dataset would allow us to use more granular sentence labels, which could potentially address the second issue. Further MTurk experiments like the experiment described in sections 3.4, 4.3, and 5.3 could yield additional insights into topical cues that arouse suspicion or trust.

Chapter 6

CONCLUSION

Product reviews provide a fascinating window into the human psyche: into relationships, insecurities, hope, gratification, disappointment, and even—one could argue—self-deception. Taking into account the apparent abundance of genuine and deceptive product reviews published online today, they can be viewed as highlighting both the essential ways in which humans engage with material objects, and the elements that advertisers perceive consumers as seeking in their purchases. Unfortunately, it is not always clear whether a particular review is an example of the former, the latter, or both, given that today's online reviewers seem to comprise a mixture of genuine consumers, hired writers, and customers motivated by promotional samples.

In this work, we have taken a variety of approaches to detecting deception in the online product review space. Drawing from the literature on deception detection, advertising, computational stylometry, and fake review detection, we derive a list of text and metadata features that we believe may correlate with deception on a product and/or review level. We calculate the correlation between various pairs of features to assess their effectiveness as gold-standard labels and/or features in binary classifiation experiments. In this process, we observe that products with particularly skewed review star-rating distributions appear to have a higher density of sentiment words in their 5-star reviews and more homogeneous language across reviews. We also observe that the density of sentiment words in 5-star reviews has increased over time, and that the variance of the creation dates of a product's reviews correlates negatively with the density of sentiment words in the reviews. Based on any one of the common assumptions that products with suspicious reviews have particularly skewed star rating distributions, that clusters of reviews created over a short period of time

indicate suspicious activity, and that the rate of suspicious reviews has increased over time, these correlations indicate sentiment words and language homogeneity as effective features for detecting products with deceptive reviews.

We explore methods of classification of products as suspicious or trustworthy (where "suspicious" is defined as "containing a high concentration of suspicious reviews"), using products' review star-rating skewedness to generate gold-standard labels. We achieve a classification accuracy of 63.03% on a held-out test set with a model that uses POS n-gram features, suggesting that products with deceptive reviews may share certain subtle stylistic elements.

Through a series of human performance experiments, we collect labels for a subsequent experiment in classifying reviews as suspicious or trustworthy, calculate human inter-annotator agreement, and gather information about the content and stylistic elements that arouse suspicion in humans. We find that humans have a fairly high level of agreement about whether or not a particular review is suspicious or trustworthy, but that these judgments do not align with the correlations found in our dataset. In general, humans rely more on reasoning about the motivations of the reviewer or the coherence of the review storyline than on the stylistic elements we found to be effective for classification.

While this initially began as a project about deception detection in the product review space, it grew into an exploration of the effectiveness of using insights from the analysis of advertising to characterize and detect fake reviews. Many of the traditional linguistic markers of deception, such as passive voice and a lack of personal pronouns, contrast with linguistic trends in advertising language. This may be in part because advertising involves a more conscious effort to dispel any fear of deception on the part of the audience, and perhaps a lesser degree of guilt. Humans appeared to use insights from both fields in detecting suspicious reviews. A lack of detail aroused suspicion, a trend which aligns with observations from the literature on deception detection. Humans were also leery of traditional advertising devices like commands, linguistic parallelism, and brand name-dropping. However, personal anecdotes tended to arouse trust, despite the fact that past research on fake review detection

and literature on advertising suggest that these anecdotes may actually be indicative of an intent to deceive and/or advertise. Unfortunately, our hand-crafted features based on insights from the literature on advertising and our human performance experiments did not show promise for the effective detection of suspicious products; we achieved optimal classification performance using shallow stylometric features such as function words and POS n-grams. This further emphasizes the observation that humans are not good at detecting deception in the product review space, and suggests that companies could benefit from machine learning based methods for flagging deceptive reviews and/or sellers engaging in suspicious activity.

One limitation of our assessment of human performance in detecting deception is that while our machine learning experiments focused on the classification of products as suspicious or trustworthy, our human performance experiments asked participants to classify reviews rather than products. Our own experience in manually reviewing product pages suggests that it is extremely difficult for a human to make a product-level judgment based on hundreds or thousands of reviews. Furthermore, experiments in product-level labeling would be difficult to deploy and monitor in a crowdsourcing setting such as MTurk. Since the text for all reviews of a deceptive product necessarily subsumes the text of many deceptive reviews, the fact that the textual features that were effective in automatically classifying products do not correlate with human review-level labels suggests that human perceptions of review deceptiveness differ dramatically from actual review deceptiveness. However, much could be learned from additional experimentation and a more systematic comparison of human and machine predictions, including human perception experiments at the product level.

While we observed a number of interesting trends suggesting the existence of clusters of products with distinctive stylistic elements that align with star-rating distribution, additional research would be needed to verify the connection between star-rating distribution and suspicious activity and, by extension, the connection between these stylistic elements and the presence of deceptive reviews. Furthermore, additional work would be needed to translate our findings into effective applications for flagging deceptive reviews and/or suspicious products.

Chapter 7

FUTURE WORK

While many interesting observations emerged from the experiments described in this work, additional research is needed to solidify the connection between deceptive reviews and the metadata features we used as labeling heuristics, primarily the skewedness of the distribution of a product's review star ratings.

In addition, while there appears to be a stylistic difference between products labeled as suspicious and trustworthy, respectively, by our heuristic, it is difficult for us to explain the particular POS and lexical features to which our classifier assigns a high weight. In future work, we would like to explore whether these features have similar predictive power for detecting deception and/or advertising intent in other domains.

Fake review detection, and deception detection in general, are particularly challenging fields in that it is difficult to develop a dataset with a reliable gold standard that is also representative of actual, unsolicited deception. Based on our results, our heuristic of using review star-rating distribution to generate product-level labels shows promise. In analyzing our dataset, we also noticed a preponderance of reviews containing the disclaimer "I received this in exchange for my honest review." These reviews had higher star ratings on average, and exhibited many elements of advertising. In future work, we would like to explore using this disclaimer to build a dataset of suspicious reviews, stripping the disclaimer from the review for the purposes of training and testing.

Our human performance experiments were somewhat limited in scope. A significant limitation is that while our machine learning experiments focused on the classification of products as suspicious or trustworthy (where "suspicious" is defined as "containing a high concentration of suspicious reviews"), our human performance experiments asked participants to

classify reviews rather than products. While our results suggest that human perceptions of review deceptiveness differ dramatically from actual review deceptiveness, much could be learned from additional experimentation and a more systematic comparison of human and machine predictions, including human perception experiments at the product level.

Our brief exploration of using discourse modeling to classify reviews as suspicious or trustworthy by human standards was not successful. However, a more thorough exploration of this approach with a significantly larger dataset would be necessary to rule it out entirely. Reviews are neatly encapsulated examples of the types of stories humans tell one another, often expressed in the casual speech that is lacking in many of the corpora traditionally used for computational linguistics research. A comparison of the results of our human performance experiments and our product-level classification experiments suggests that humans' perceptions of deceptive reviews may not align with actual characteristics of deceptive reviews, so building statistical models of reviews' trustworthiness by human standards may not be a fruitful direction from a fake review detection standpoint. However, this approach could contribute to our understanding of human's perceptions of trustworthiness, perhaps beyond the domain of product reviews.

BIBLIOGRAPHY

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. Success with style: Using writing style to predict the success of novels. *Poetry*, 580(9):70, 2013.

Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. *arXiv preprint cs/0405039*, 2004.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

Guy Cook. *The discourse of advertising*. Psychology Press, 2001.

Rakefet Dilmon. Between thinking and speaking—linguistic tools for detecting a fabrication. *Journal of Pragmatics*, 41(6):1152–1170, 2009.

Song Feng, Ritwik Banerjee, and Yejin Choi. 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, 2012a.

Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. Distributional footprints of deceptive product reviews. In *ICWSM*, 2012b.

Pedro A Fuertes-Olivera, Marisol Velasco-Sacristán, Ascensión Arribas-Baño, and Eva Samaniego-Fernández. Persuasion and advertising english: Metadiscourse in slogans and headlines. *Journal of Pragmatics*, 33(8):1291–1307, 2001.

Sabine Gieszinger. *The History of Advertising Language: The Advertisements in The Times from 1788 to 1996*. Peter Lang: Europäischer Verlag der Wissenschaften, 2001.

Stephanie Gokhman, Jeff Hancock, Poornima Prabhu, Myle Ott, and Claire Cardie. In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 23–30. Association for Computational Linguistics, 2012.

Jeffrey T Hancock, Jennifer Thom-Santelli, and Thompson Ritchie. Deception and design: The impact of communication technology on lying behavior. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 129–134. ACM, 2004.

Manoj Harpalani, Michael Hart, Sandesh Singh, Rob Johnson, and Yejin Choi. Language of vandalism: Improving Wikipedia vandalism detection via stylometric analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers—Volume 2*, pages 83–88. Association for Computational Linguistics, 2011.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 219–230. ACM, 2008.

Geoffrey N Leech. *English in advertising: A linguistic study of advertising in Great Britain*. Longmans, 1966.

Lemur Project. URL `http://www.lemurproject.org/`.

Jiwei Li, Claire Cardie, and Sujian Li. TopicSpam: a topic-model based approach for spam detection. In *ACL (2)*, pages 217–221, 2013a.

Jiwei Li, Myle Ott, and Claire Cardie. Identifying manipulated offerings on review portals. In *EMNLP*, pages 1933–1942, 2013b.

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. Towards a general rule for identifying deceptive opinion spam. ACL, 2014.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit, 2002. URL `http://mallet.cs.umass.edu`.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. Structured models for fine-to-coarse sentiment analysis. In *Annual Meeting-Association for Computational Linguistics*, volume 45, page 432. Citeseer, 2007.

Rohith Menon and Yejin Choi. Domain independent authorship attribution without domain adaptation. In *RANLP*, pages 309–315. Citeseer, 2011.

Rada Mihalcea and Carlo Strapparava. 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, 2009.

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

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

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 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, 2011.

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

Myle Ott, Claire Cardie, and Jeffrey T Hancock. Negative deceptive opinion spam. In *HLT-NAACL*, pages 497–501, 2013.

James W Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

Dinesh Puranam, Samuel Curtis Johnson, and Claire Cardie. The enrollment effect: A study of Amazon's Vine program. *ACL 2014*, page 17, 2014.

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

Torben Vestergaard and Kim Schrøder. *The language of advertising*. Blackwell Oxford, 1985.

Aldert Vrij. *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons, 2008.

Aldert Vrij, Sharon Leal, Pär Anders Granhag, Samantha Mann, Ronald P Fisher, Jackie Hillman, and Kathryn Sperry. Outsmarting the liars: The benefit of asking unanticipated questions. *Law and human behavior*, 33(2):159–166, 2009.

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

Lina Zhou, Judee K Burgoon, Jay F Nunamaker, and Doug Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, 13(1):81–106, 2004.

Lina Zhou, Yongmei Shi, and Dongsong Zhang. A statistical language modeling approach to online deception detection. *Knowledge and Data Engineering, IEEE Transactions on*, 20(8):1077–1081, 2008.