Opinion Spam and Analysis

NITIN JINDAL & BING LIU, WSDM 08

UIUC

Opinion/Review Spam

All spam is spam but some spam is more spam than others

Opinion spam similar to web spam or email spam in intent, but different in form / content

- Web spam – uses illegitimate means to boost web page rank in search engines
- Email spam – has advertising or indiscriminate delivery of unsolicited content

50 Shades of Opinion Spam

Just three, actually:

Type 1: Untruthful Opinions

- Very virulent kind of spam
- Deliberately mislead readers or automated systems by giving false positive or false negative reviews
- Called Fake reviews / Bogus reviews

Type 2: Reviews on brands only

- Do not contain specific product reviews but rather just reviews for brands / manufacturers / sellers
- May be useful; treated as spam in present study

Type 3: Non-reviews

- Non-reviews, such as ads, or other irrelevant text without opinions

Key Issues

General Spam Detection is treated as a classification problem where the classes are simply {SPAM, NOT SPAM}

This works well for Type 2 (non-specific reviews) and Type 3 (non-reviews) spam

Manual labeling of Type 1 Spam very difficult

WHY?

Dataset

Reviews scraped from Amazon.com

- 5.8m reviews
- 2.14m reviewers
- 6.7m products

Fields for each review: Product ID, Reviewer ID, Rating, Date, Review Title, Review Body, Number of Helpful Feedbacks, Number of Feedbacks

Observations

- Number of reviews v/s Number of reviewers follow a power law distribution
- Quite a few 'similar reviews': More on that later

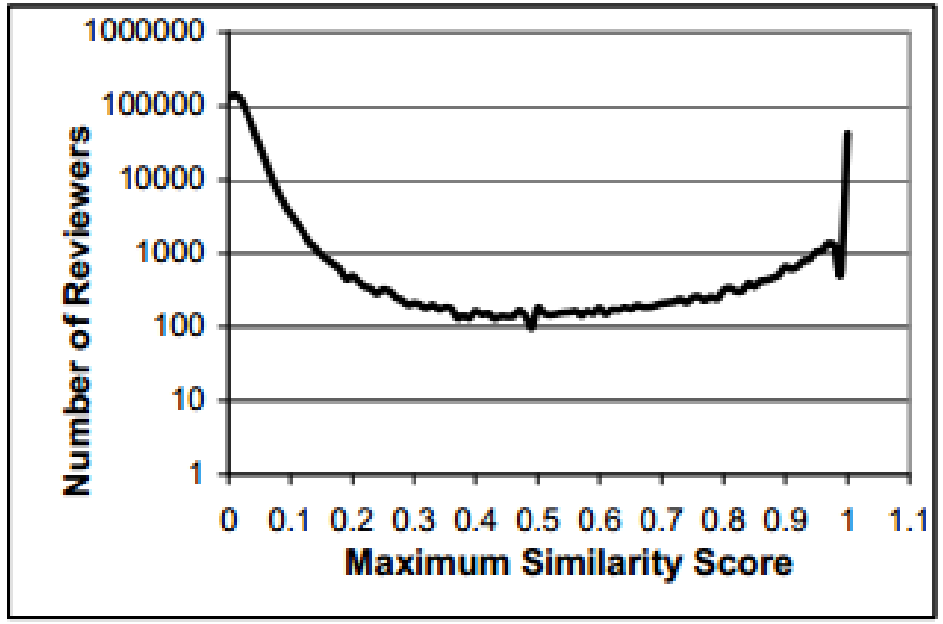
Duplicates Duplicates Everywhere Everywhere!

Three kinds of iffy duplicates

- Different user-ids on same product
- Same user-id on different product
- Different user-id on the different products

Duplicates detected using Jaccard Distance

- N(**A** AND **B**) / N(**A** OR **B**)
- 2-gram features



You know it, but can you prove it?

Spam types 2 and 3 are easy to classify manually; yay labeled data!

Use logistic regression and see if it can reliably identify Type 2 and Type 3 Spam

36 features:

- Review Centric: Feedback, Length, Position, % of +ve and -ve opinion words, similarity with product features, % of numerals, capital letters etc. yadda yadda
- Reviewer Centric: **Guess?**
- Product Centric: Price, Sales rank, Rating, Deviation in rating etc.

Spam Type	Num reviews	AUC	AUC – text features only	AUC – w/o feedbacks
Types 2 & 3	470	98.7%	90%	98%
Type 2 only	221	98.5%	88%	98%
Type 3 only	249	99.0%	92%	98%

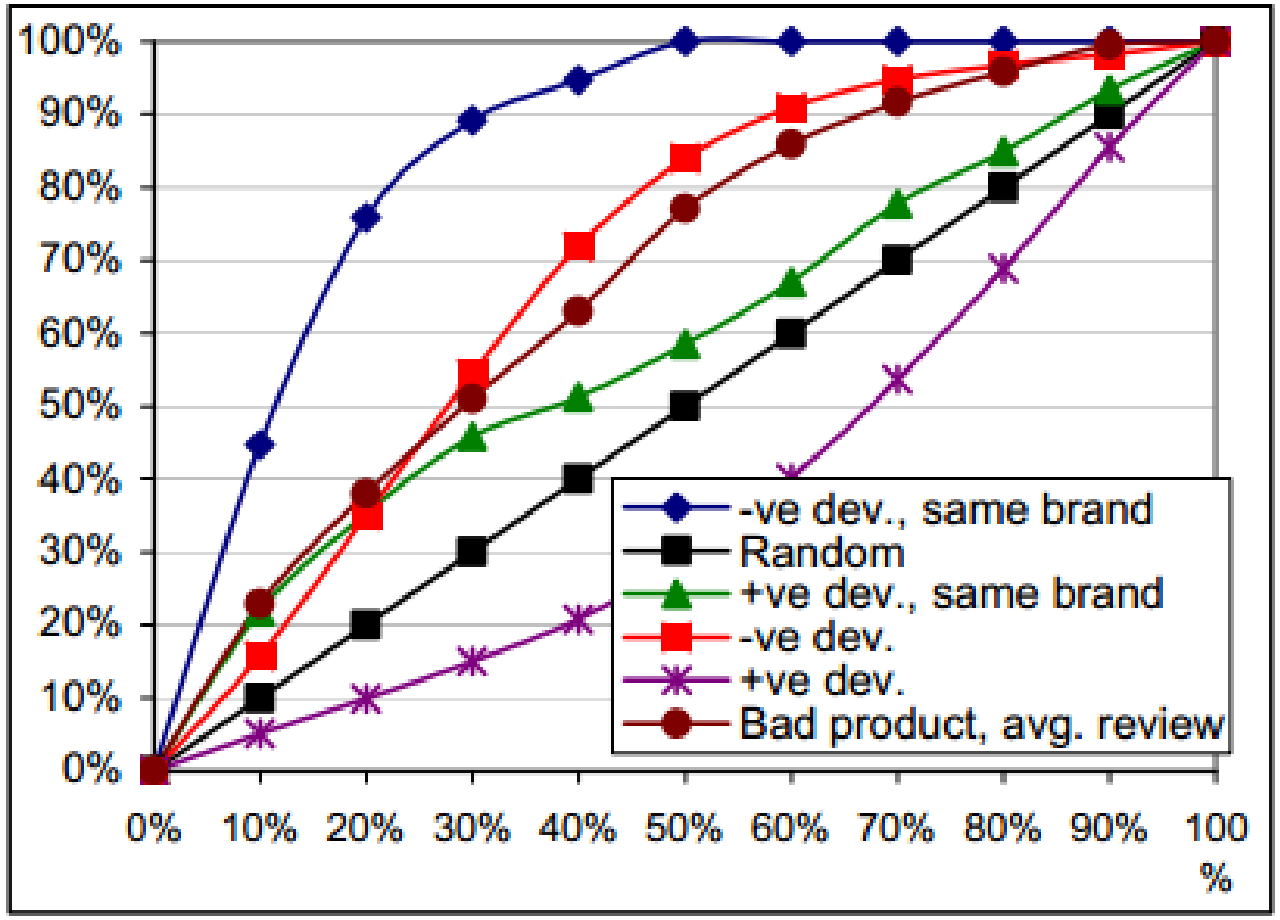
Yeah, fine but what about Type 1?

Treat duplicate reviews as SPAM to see if they can be predicted

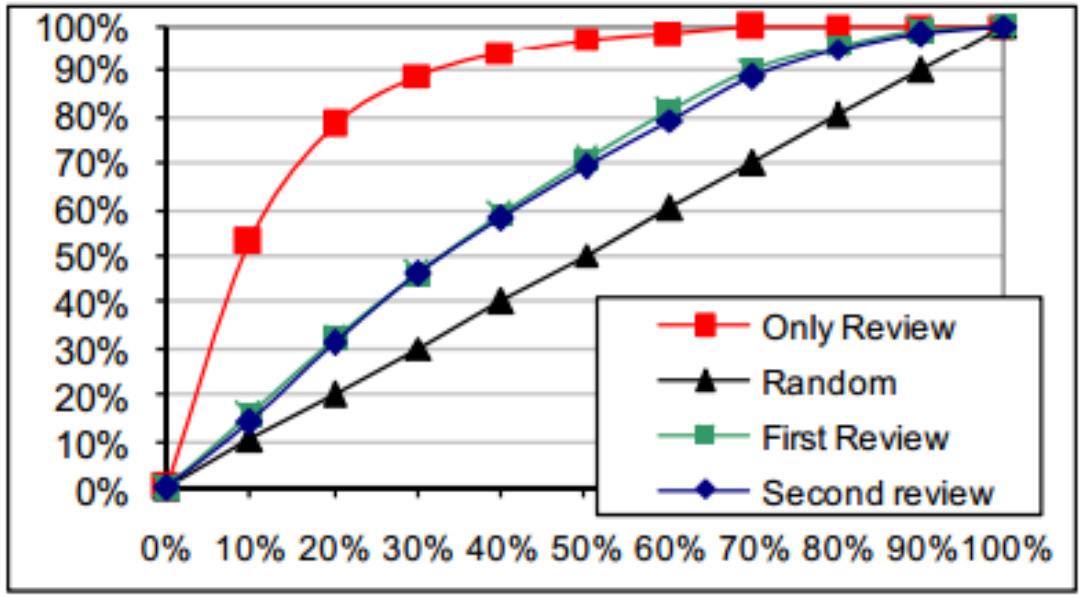
Features used	AUC
All features	78%
Only review features	75%
Only reviewer features	72.5%
Without feedback features	77%
Only text features	63%

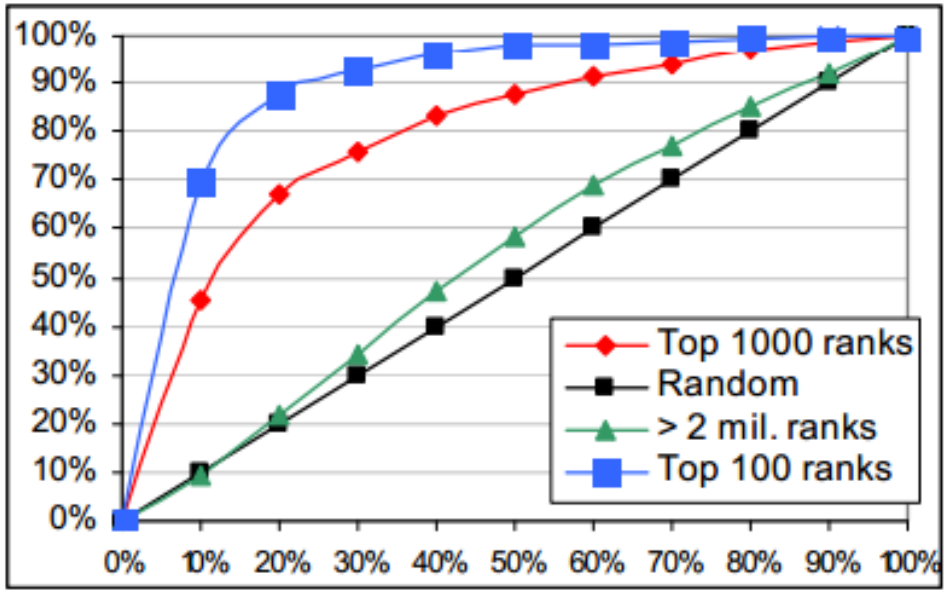
Try to predict *outlier reviews* (whose rating goes against the grain of the overall rating)

Lift Curves


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100
%
-ve dev., same brand
Random
+ve dev., same brand
-ve dev.
+ve dev.
Bad product, avg. review

Lift Curves


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Only Review
Random
First Review
Second review


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Top 1000 ranks
Random
> 2 mil. ranks
Top 100 ranks

Discussion

Lots of interesting results

Sets a good baseline and 'ground terms' for future work

Some of the explanations for the curves seem a bit 'hand-wavy'

Distortion as a Validation Criterion in the Identification of Suspicious Reviews

GUANGYU WU, DEREK GREENE, BARRY SMYTH, PADRAIG CUNNINGHAM

SOMA 2010

Motivation

Type 1 Opinion Spam

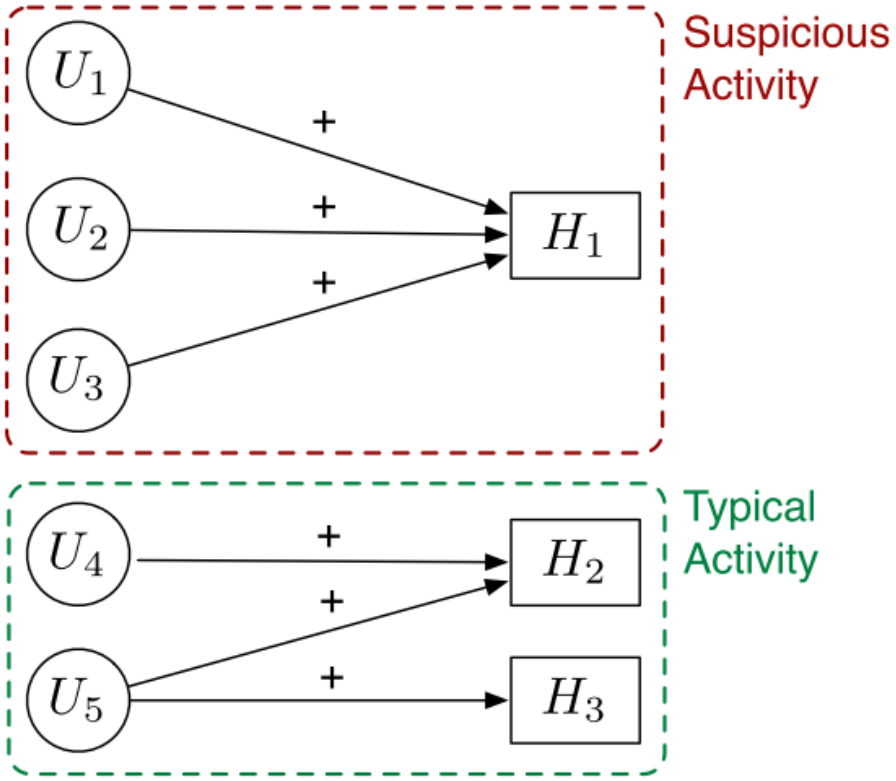
Automatically detect a subset Type 1 opinion spam, false-positive "shill" reviews.

Hotel Review Dataset

TripAdvisor.com

29,799 reviews; 21,851 unique reviewers; hotels in Ireland over a 2-year period

Positive Singleton Detection



Measures

Proportion of Postive Singletons $PPS(H) = \frac{N_{ps}}{N}$

Concentration of Positive Singletons

$$CPS(H) = \frac{1}{P}\sum_{i=1}^{P} e^{-\lambda \times min(D(r_i,r_{i-1}),D(r_i,r_{i+1}))}$$

Distortion

Raw Distortion

Spearman rank correlation.

Adjusted Distortion

Normalizing distortion on number of reviews.

Significant adjusted distortion scores will be positive.

Insignificant adjusted distortion scores will be close to zero.

Distortion

Raw Distortion

Spearman rank correlation.

Adjusted Distortion

Normalizing distortion on number of reviews.

Significant adjusted distortion scores will be positive.

Insignificant adjusted distortion scores will be close to zero.

Results on TripAdvisor Dataset

Nothing!

Talked about one hotel that had suspicious reviews, but then dismissed them on the basis of, "we looked at the reviews and they seemed legit".

Didn't actually provide or discuss results because they couldn't be validated.

"We plan to explore this issue in further work."

Finding Deceptive Opinion Spam by Any Stretch of the Imagination

MYLE OTT, YEJIN CHOI, CLAIRE CARDIE, JEFFREY T. HANCOCK

CORNELL UNIVERSITY, 2011

Motivation

Disruptive opinion spam

Uncontroversial instances of spam that are easily identified by a human reader.

Deceptive opinion spam

Fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader.

Which one is spam?

I was apprehensive after reading some of the more negative reviews of the Hotel Allegro. However, our stay there was without problems and the staff could not have been more friendly and helpful. The room was not huge but there was plenty of room to move around without bumping into one another. The bathroom was small but well appointed. Overall, it was a clean and interestingly decorated room and we were pleased. Others have complained about being able to clearly hear people in adjacent rooms but we must have lucked out in that way and did not experience that although we could occasionaly hear people talking in the hallway. One other reviewer complained rather bitterly about the area and said that it was dangerous and I can't even begin to understand that as we found it to be extremely safe. The area is also very close to public transportation (we used the trains exclusively) and got around quite well without a car. We would most definetly stay here again and recommend it to others.

I went here with the family, including our dog Marley(They are very pet friendly). We really enjoyed it. This place is huge with over 480 rooms and suites and is in the center of downtown close to shopping and entertainment. It also seems that it would be a great place to have a wedding or to host an event. I will definately be coming back next time I need to come to chicago definately a fine four star hotel!

Which one is spam?

I was apprehensive after reading some of the more negative reviews of the Hotel Allegro. However, our stay there was without problems and the staff could not have been more friendly and helpful. The room was not huge but there was plenty of room to move around without bumping into one another. The bathroom was small but well appointed. Overall, it was a clean and interestingly decorated room and we were pleased. Others have complained about being able to clearly hear people in adjacent rooms but we must have lucked out in that way and did not experience that although we could occasionaly hear people talking in the hallway. One other reviewer complained rather bitterly about the area and said that it was dangerous and I can't even begin to understand that as we found it to be extremely safe. The area is also very close to public transportation (we used the trains exclusively) and got around quite well without a car. We would most definetely stay here again and recommend it to others.

I went here with the family, including our dog Marley(They are very pet friendly). We really enjoyed it. This place is huge with over 480 rooms and suites and is in the center of downtown close to shopping and entertainment. It also seems that it would be a great place to have a wedding or to host an event. I will definately be coming back next time I need to come to chicago definately a fine four star hotel!

Which one is spam?

My husband and I satayed for two nights at the Hilton Chicago,and enjoyed every minute of it! The bedrooms are immaculate,and the linnens are very soft. We also appreciated the free wifi,as we could stay in touch with friends while staying in Chicago. The bathroom was quite spacious,and I loved the smell of the shampoo they provided-not like most hotel shampoos. Their service was amazing,and we absolutely loved the beautiful indoor pool. I would recommend staying here to anyone.

Thirty years ago, we had a tiny "room" and indifferent service. This time, the service was superb and friendly throughout, with special commendation for the waiters and waitresses at the coffee shop, the door and bell persons, and the hilton honors person at the front desk. They even lowered our price (to moderately high) when we inquired a few days before our stay. When we want to stay south of the river downtown, we will be back

Which one is spam?

My husband and I satayed for two nights at the Hilton Chicago,and enjoyed every minute of it! The bedrooms are immaculate,and the linnens are very soft. We also appreciated the free wifi,as we could stay in touch with friends while staying in Chicago. The bathroom was quite spacious,and I loved the smell of the shampoo they provided-not like most hotel shampoos. Their service was amazing,and we absolutely loved the beautiful indoor pool. I would recommend staying here to anyone.

Thirty years ago, we had a tiny "room" and indifferent service. This time, the service was superb and friendly throughout, with special commendation for the waiters and waitresses at the coffee shop, the door and bell persons, and the hilton honors person at the front desk. They even lowered our price (to moderately high) when we inquired a few days before our stay. When we want to stay south of the river downtown, we will be back

Goals

1. Create a gold-standard opinion spam dataset.

2. Develop and compare three approaches to detectiving deceptive opinion spam.

    - Genre classification
    - Psycholinguistic deception detection
    - Text categorization

Task 1: Creating the corpus

Truthful reviews are taken from TripAdvisor.com (5 stars only)

Deceptive reviews are created by Mechanical Turkers (positive reviews only)

20 truthful and 20 deceptive reviews for each of 20 hotels

800 reviews total

Task 2: Human performance

Why?

Need a baseline to analyze automatic methods against.

If human performance is low, then the importance of the task increases.

How?

Mechanical Turk didn't work -- used three undergrads instead.

Meta-judge (majority and skeptic)

Task 2: Human performance results

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

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

Task 2: Three automated approaches

1. Genre identification
2. Psycholinguistic deception detection
3. Text categorization

Task 2: Genre identification

POS tags as features

Why?

Frequency distribution of POS tags has been shown to be dependent on the genre of the text

Task 2: Psycholinguistic deception detection

Linguistic Inquiry and Work Count (**LIWC**) software

- LIWC counts and groups 4500 keywords into 80 psychologically meaningful dimensions

Method: create classifier using the LIWC dimensions as features for the classifier

Task 2: Psycholinguistic deception detection

The 80 LIWC features can be summarized into four categories:

1. Linguistic processes (e.g. average # words per sentence)
2. Psychological processes (e.g. social, emotional, cognitive, perceptual, time, space...)
3. Personal concerns (e.g. work, leisure, money, religion...)
4. Spoken categories (e.g. filler and agreement words)

Task 2: Text categorization

- unigrams
- bigrams
- trigrams

Task 2: Classifiers for the 3 methods

1. Naive Bayes
2. Linear Support Vector Machine (SVM)

---> Test each of the 3 methods and also all combinations of methods.

Task 2: Results

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

Task 2: Results

- Computers beat humans on every metric
  - Except for Judge 2 who has uncannily good truthful recall
- Untrained humans often focus on unreliable cues to deception (e.g. second-person pronouns)
- The genre classifier beats humans too!
  - Maybe truth/deceptive correlates with informative/imaginative genres
- Best performance = LIWC + Bigrams

Task 2: Results

- Truthful opinions have more sensorial and concrete language
- Truthful opinions are more specific about spatial configurations
- Deceptive opinions focus on aspects external to the hotel
  - e.g. husband, business, vacation
- Deceptive reviews have more positive and fewer negative emotion terms
- Deceptive reviews use more first person singular

Discussion

- Only very positive reviews were studied
- Do the results mean that we should we throw away the psychology?
- Humans vs. computers (if the humans were trained, would they still perform as poorly?)
- Why do we care about deceptive reviews?

Read more here:

http://www.nytimes.com/2012/08/26/business/book-reviewers-for-hire-meet-a-demand-for-online-raves.html?pagewanted=all&_r=0