Finding Deceptive Opinion Spam by Any Stretch of the Imagination

Authors: Myle Ott, Yejin Choi, Claire Cardie, Jeffrey T. Hancock

Presenter: Yong Xie
University of Illinois

CONTENT

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

- Background and Motivation
- Dataset Construction
- Automated Classification

BACKGROUND

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

- Deceptive opinion spam: fictitious opinions that have been deliberately written to sound authentic.

I have stayed at many hotels traveling for both business and pleasure and I can honestly stay that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to both business travellers and couples.

My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The area of the hotel is great, since I love to shop I couldn't ask for more!! We will definitely be back to Chicago and we will for sure be back to the James Chicago.

Examples of truthful and deceptive opinions. *Source: OCCH 2011*

- Challenges:
  - Deceptive opinion spam are insidious and stealthy
  - Few good sources of labeled data

DATA CONSTRUCTION - SOURCE

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

- Dataset of 800 opinions: 400 truthful and 400 gold-standard deceptive reviews.
- Truthful opinions from 5-star reviews from 20 hotels on TripAdvisor.
- Deceptive opinions via Amazon Mechanical Turk

DATA CONSTRUCTION – HUMAN EVALUATION

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

- Three student judges without monetary reward.

- Two virtual meta-judges
  - Majority meta-judge: predicts deceptive when at least two out of three believe the review to be deceptive.
  - Skeptical meta-judge: predicts deceptive when any human judge believes the review to be deceptive.

DATA CONSTRUCTION – HUMAN EVALUATION

			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**

Results of human evaluation; *Source: OCCH 2011*

- Truthful opinions have high recall but low precision, whereas deceptive opinions have low recall but high precision.
- More positive prediction leads to higher recall but lower precision, so human tend to classify as truthful.
- The overall accuracy is just slightly higher than random guess.

AUTOMATIC METHODS – GENRE IDENTIFICATION

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

- Part-of-speech (POS) tag as features, and SVM as classifier.
- The weight parameters demonstrate the importance of POS tags to each category.
- Informative writing: more nouns, adjectives, prepositions, determiners, and coordinating conjunctions.
- Imaginative writing: more verbs, adverbs, pronouns, and pre-determiners.

AM – PSYCHOLINGUISTIC APPROACH

- Linguistic Inquiry and Word Count (LIWC): counts and groups the number of instances of nearly 4,500 keywords into 80 psychologically meaningful dimension.

- LIWC output as features, SVM as classifier.

AM – TEXT CATEGORIZATION

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

- Two classifiers: linear SVM and Naïve Bayes.
- Feature set
  - Unigrams, bigrams, trigrams, smoothed by interpolated Kneser-Ney method
  - LIWC
- Training scheme: 5-fold cross validation

AM – 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

Model performance of automated classifiers. *Source: OCCH 2011*

- Automatic classifiers outperform human judges for most metric.
- SVM trained on bigram + LIWC achieves the best performance overall.
- Automatic classifiers are more consistent and balanced than human judges.
- Context improves mode performance.
- SVM outperforms NB.

AM – INTERPRETABILITY

TRUTHFUL/INFORMATIVE			DECEPTIVE/IMAGINATIVE		
Category	**Variant**	**Weight**	**Category**	**Variant**	**Weight**
NOUNS	Singular	0.008	VERBS	Base	-0.057
	Plural	0.002		Past tense	**0.041**
	Proper, singular	**-0.041**		Present participle	-0.089
	Proper, plural	0.091		Singular, present	-0.031
ADJECTIVES	General	0.002		Third person singular, present	**0.026**
	Comparative	0.058			
	Superlative	**-0.164**		Modal	-0.063
PREPOSITIONS	General	0.064	ADVERBS	General	**0.001**
DETERMINERS	General	0.009		Comparative	-0.035
COORD. CONJ.	General	0.094	PRONOUNS	Personal	-0.098
VERBS	Past participle	0.053		Possessive	-0.303
ADVERBS	Superlative	**-0.094**	PRE-DETERMINERS	General	**0.017**

Average feature weights learned by SVM on POS features. *Source: OCCH 2011*

- Weights learned by SVM on POS tag features are in line with distinction between imaginative and informative writings.
- How does the results generalize to other dataset of deceptive reviews?

AM – INTERPRETABILITY

LIWC+BIGRAMS$^{+}_{SVM}$		LIWC$_{SVM}$	
TRUTHFUL	DECEPTIVE	TRUTHFUL	DECEPTIVE
-	chicago	hear	i
...	my	number	family
on	hotel	allpunct	perspron
location	,_and	negemo	see
)	luxury	dash	pronoun
allpunct$_{LIWC}$	experience	exclusive	leisure
floor	hilton	we	exclampunct
(	business	sexual	sixletters
the_hotel	vacation	period	posemo
bathroom	i	otherpunct	comma
small	spa	space	cause
helpful	looking	human	auxverb
$	while	past	future
hotel_.	husband	inhibition	perceptual
other	my_husband	assent	feel

Top 15 highest weighted truthful and deceptive features

Source: OCCH 2011

- Truthful opinions tend to include more sensorial and concrete language.
- Truthful opinions are more specific about spatial configuration.
- Increased focus in deceptive opinions on aspects external to the hotel being reviewed

COMMENT

- It seems that the fundamental difference between deceptive and truthful reviews is that they describe different aspects of the hotels.
- The difference is reflected on the vocabulary as well, that is the reason why unigram feature works very well.
- Context helps classification performance, but the improvement is limited. Does better language model help?
- The dataset is very limited, in terms of size and diversity. If human annotation is unavailable, can we resort to unsupervised or semi-supervised method to train classifers?

THANK YOU

Q & A