CAPTURING WORD CHOICE PATTERNS WITH LDA FOR FAKE REVIEW DETECTION IN SENTIMENT ANALYSIS

KYUNGYUP DANIEL LEE

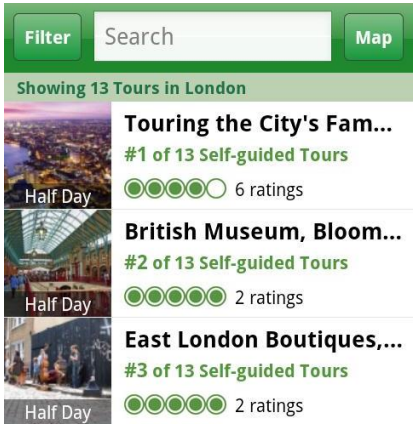
DUKIN CO., LTD.

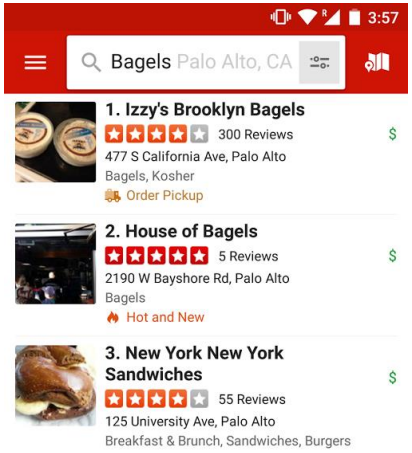
LKYLOVE2323@GMAIL.COM

CONTENTS

INTRODUCTION

- **User-generated online reviews are becoming increasingly important for potential customers in making purchase decisions (e-tailing group and PowerReviews, 2011)**
- **Businesses like hotels or restaurants have attempted to hire people to generate deceptive reviews**
  - Praise their business
  - Depreciate others





INTRODUCTION

- **In this research, we defined it as follow**

> Fake review: a factitious one written by a person with no or little experience in the project or service, who has been asked to lavishly praise it

INTRODUCTION

- **Detecting fake review by its text contents is not easy task.**
  - Human has low performance for detecting deception (Mark A. 1991, Fiedler and Walka, 1993)
  - Detecting fake review by utilizing only text contents show lower performance than by utilizing meta data (Mukherjee et al., 2013)
- **Research about fake review detection is challenging work.**
  - Due to the difficulty of collecting labeled data.

CHARACTERIZING FAKE REVIEWS

Unnatural Word Choice

- **We assumed that writing a fake review is a type of telling a lie (Shozo, 2004).**
  - **Self-awareness**: Fake reviewers know that what they are writing is likely to be different from the fact.
  - **Design to deceive**: There is an intention to convince people as if it is truthful using plausible expressions.
  - **Specific aim**: Telling a lie has a clear purpose.
- **Telling a lie is subject to an uncomfortable word choice process (Meyer, 2011)**
  - **Paralinguistic behaviors**: speech rate, intonation, accent, etc.
- **So writing a fake review can be subject to unnatural word choice.**

CHARACTERIZING FAKE REVIEWS

Word Choice and Topics

- **Word choice in linguistics is a post-process referred to as conceptualization (Levelt, 1989).**
- **The word choice process when producing text is captured well as a generative process in a computational model like LDA.**

Generative process of LDA

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\boldsymbol{\theta} \sim \text{Dir}(\alpha)$.
3. For each of the N words
   A. Choose a topic $z_n \sim \text{Multinomial}(\boldsymbol{\theta})$.
   B. Choose a word w_n from $p(w_n|z_n, \boldsymbol{\beta})$, a multinomial probability conditioned on the topic z_n.

Figure 1: The generative process of LDA

CHARACTERIZING FAKE REVIEWS

Word Choice and Topics

- **We hypothesize:**
  - The “topics” from which words are drawn in fake and truthful reviews are likely to be different from each other and hence captured accordingly.

DATASET FOR VALIDATION

- **We developed our own dataset that includes seven business domains.**
- **It is based on reviews on the Yelp site.**
  - Only commercial review site that includes the labeling of fake reviews
  - Yelp has its own filtering system which have utilized a variety of large-scale data and behavioral meta-features since 2005 (Yelp, Inc., 2009) .

DATASET FOR VALIDATION

- **To ensure that the dataset we developed reflects the real world, we first crawled all the existing reviews belonging to the seven domain categories in the Chicago area from Yelp and then filtered out those not satisfying some requirements (May, 2014).**
- **We also employed under-sampling as in past work for fair comparisons (Ott et al., 2011, Mukherjee et al., 2013).**

					After filtering	
Category	# of business	# of truthful	# of fake	% of fake	# of truthful	# of fake
Electronics	34	1,027	973	48.7%	220	220
Fashion	394	12,920	5,370	29.4%	1,345	1,345
Hospital	27	884	652	42.4%	100	100
Hotel	136	6,433	2,757	30.0%	550	550
Insurance	5	129	142	52.4%	65	65
Music venue	94	5,463	4,040	42.5%	595	595
Restaurant	996	418,430	60,757	12.7%	37,980	37,980

FIVE TOPIC TYPES

- **To grasp the traits of "topics" for truthful and fake reviews, we attempted to group the extracted "topics" from LDA into five types we suggested**
  - **Concrete Experience (CE)**: Topics in this group mainly contain verbs.
  - **Detailed Information (DI):** Topics of this type are usually composed of specific nouns and adjectives.
  - **General Comments (GC):** Topics in this class include words describing abstract evaluations.
  - **Comparative Assessment (CA):** Topics of this nature usually include comparative words.
  - **Recommendation and Reference (RR):** Topics of this type tend to include words like "recommend", "refer", etc.

FIVE TOPIC TYPES

- **Here are example reviews with topic words with its type.**

A truthful review	A fake review
I try to go here to find great deals on work clothes. I try to buy all my pants for work here, but then have always found some other great finds. … I could never pay full price for … I would recommend this place to anyone who like to save money	Great place to find unique outfits to go out and work… Prices are also lower than other stores in the Wickerpark area and the quality is surprisingly better than Zara, Guess … I came in one day because a dress caught my eye and I was wearing winter boots that day :(

Figure 2: Examples of a truthful and a fake review. The Color of words denotes type of topic which the word is included in: CE (blue), DI (green), GC (yellow), CA (red) and RR (orange).

T/F TENDENCY OF TOPIC TYPES

- We first generated 100 topics by applying LDA to a collection of reviews.
- Every topic was then classified into one of the five topic types or set aside as belonging to the “none of the above”
- The elements of a vector resulting from the SVM classifier with a linear kernel have an indication for each class about relevance and direction

T/F TENDENCY OF TOPIC TYPES

- We found that the average weight of each topic type learned by SVM was in good agreement with our hypothesis.

Predicted to be Truthful by Humans		Predicted to be Fake by Humans	
Type of topic	Avg. Weight	**Type of topic**	Avg. Weight
Concrete Experience (CE)	**0.068**	General Comment (GC)	**- 0.148**
Detailed Information (DI)	**0.142**	Comparative Assessment (CA)	**- 0.204**
		Recommendation and Reference (RR)	**- 0.054**

Average feature weights trained by SVM-TD for five topic types

FAKE REVIEW DETECTION METHOD

- **We propose an automatic fake review classification method: Weighted topic distribution (Weighted TD)**
- **To utilize the topics biased towards one of the two groups of types, T-Type and F-Type, we devise the notion of relative global topic distributions of truthful and fake reviews, Θ_T and Θ_F.**
- **They are computed from the LDA results for the entire training document collection, consisting of two sub-collections,D_T and D_F.**

$$\Theta_X = \frac{\sum_{d \in D_X} \theta_d}{\sum_{d \in D_T \cup D_F} \theta_d}$$

where $X = \{T, F\}$

FAKE REVIEW DETECTION METHOD

- **The factors w_T^i and w_F^i are weights of the *i*-th topics in $\boldsymbol{D}_T$ and $\boldsymbol{D}_F$, respectively. The weights reflect how widely a topic is spread across the reviews in either sub-collection.**

$$\mathrm{w}_X^i = \log \frac{\left|\left\{\mathrm{d} \in \boldsymbol{D}_X : \theta_{\mathrm{d}}^i > \tau\right\}\right| + 1}{\left|\left\{\mathrm{d} \in \boldsymbol{D}_Y : \theta_{\mathrm{d}}^i > \tau\right\}\right| + 1}$$

where $(X, Y) = (T, F)$ or (F, T) , $\tau = 1/K, K \; is \; number \; of \; topics \; in \; a \; LDA \; model$

- **Finally, we combine the two factors for the final score:**

$$score_X(\mathrm{r}) = \sum_{\mathrm{K}} \theta_{\mathrm{r}}^{\mathrm{i}} \times \left(\Theta_{\mathrm{X}}^{\mathrm{i}} + \sigma \times \mathrm{w}_{\mathrm{X}}^{\mathrm{i}}\right)$$

where $X = \{T, F\}$ and $\boldsymbol{\theta}_r$ is topic distribution of test review r

FAKE REVIEW DETECTION METHOD

- **Finally, we combine the two factors for the final score:**

$$Score_X(\mathrm{r}) = \sum_{\mathrm{K}} \theta_{\mathrm{r}}^{\mathrm{i}} \times \left(\Theta_{\mathrm{X}}^{\mathrm{i}} + \sigma \times \mathrm{w}_{\mathrm{X}}^{\mathrm{i}}\right)$$

- **A review r is classified as fake review if its score for fake is greater than its score for truthful**

$$score_F(r) > score_T(r)$$

EXPERIMENTS AND ANALYSIS

Human Performance

- **Our goal in human participation study**
  - To find out whether learning would change the performance
  - To confirm the difficulty in detecting fake reviews
  - To make a fair comparison between human and automatic methods
- **Consequently 14 participants fluent in English were asked to do the following:**

1) Classification: 84 review documents in seven categories or 12 reviews for each category (1 out of 5)
2) Learning: 336 labeled reviews (4 out of 5)
3) Classification: 84 additional reviews (1 out of 5)
4) Interview: Questionnaire

EXPERIMENTS AND ANALYSIS

Human Performance

- The effect of learning was up to the participants and categories.
- Overall slightly negative effect from learning (-3.2%)
- Eight people who thought the training was helpful ended up having an improvement of 2.7%

EXPERIMENTS AND ANALYSIS

Classification Results

- Proposed model outperformed SVM-UB by 10.9% in small size categories
- But SVM-UB outperformed Weighted TD by 12.4% in large size categories

	Avg. # of reviews	SVM-U	SVM-UB	SVM-TD	Weighted TD	Human Judges
Small size categories	1,224	51.2%	51.4%	52.3%	**57.0% (+10.9%)**	48.0%
Large size categories	40,670	57.2%	**57.9%**	51.1%	**51.5% (-11.1%)**	48.2%

Accuracy of classifiers for the Yelp-based data

EXPERIMENTS AND ANALYSIS

- Larger dataset need to have large space of topics to capture the topics trend for fake review detection
  - We fixed total number of topics as 100 and it may fitted to the small data categories
- Some negligible words in n-gram classifier are distinct feature in our topic based model
  - The word ‘area’ used in fake review is belong to comparative assessment type
    - ‘*the restaurant is best in this area*’
  - But that used in truthful review is belong to concrete experience type
    - ‘*just happened to be in the area*’

Thank you!

T/F TENDENCY OF TOPIC TYPES

	Topic Type	Descriptive Label	Topic Words
T-types	CE	Purchase of computer	laptop, computer, buy, brought, refurbish, …
	DI	Repair service	equipment, gear, instrument, preamp, design, retrofit, …
F-types	GC	Overall good place	very, good, perfect, place, overall, atmosphere, …
	CA	Fabulous shop	shop, fabulous, more, s uperb, friendliest, competitive …
	RR	Recommendation	recommend, great, highly, satisfied, refer, ...

EXPERIMENTS AND ANALYSIS

Human Performance

Answer for "Was training helpful?"	Before learning	After learning	Rate of change
"Helpful."	49.6%	50.9%	**2.7%**
"So so."	49.2%	43.2%	**-12.1%**
"Not helpful."	51.2%	46.0%	**-10.1%**
Avg.	49.8%	48.2%	**-3.2%**

Table 4: Performance of the human participants