Fake Review Detection on Yelp

Tianpei Qian, Zehui Wang, Yuzhu Zhang

(tianpei/wzehui/arielzyz)@stanford.edu

Stanford University

Stanford ENGINEERING

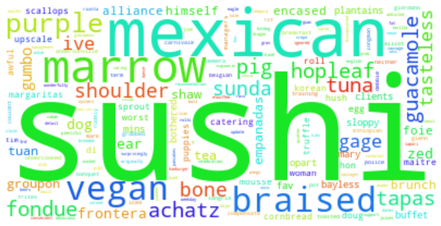
PROBLEM DESCRIPTION

- Researches has questioned the reliability of online review.
- Around 20% reviews on Yelp are estimated to be faked by paid human writers, which is misleading when consumers are making decisions.
- In this project our goal is to build a classifier to distinguish fake reviews from truthful reviews.

DATA AND FEATURE EXTRACTION

1. Dataset

The dataset contains around 8141 trustworthy and 8141 deceptive reviews for Chicago restaurants on Yelp.[1]

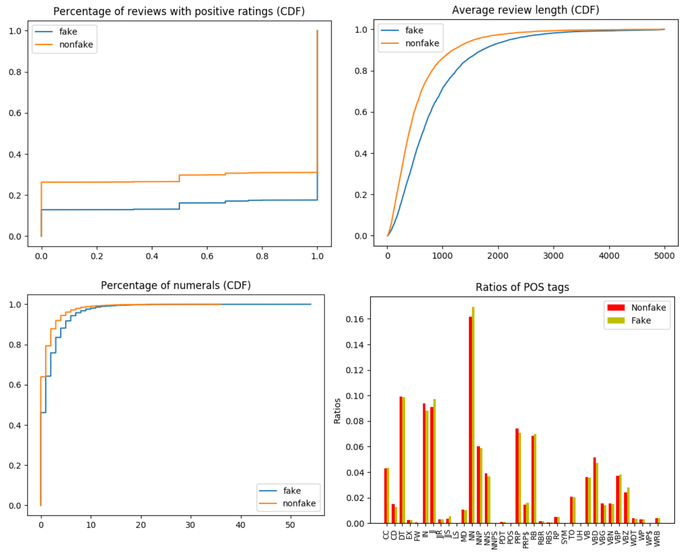


Words with the largest difference in frequency between fake and non-fake reviews

2. Feature Extraction

- **Reviewer-centric Features (5)**
  - Average review length of one reviewer (1)
  - Maximum number of reviews written in one day (1)
  - Percentage of positive/negative rating reviews (2)
  - Standard deviation of ratings (1)
- **Review-centric Features (242)**
  - Structural features (42): length, average word length, number of sentences, average sentence length, number of numerals, number of capitalized words, POS percentages
  - Unigrams & bigrams (200): With unigrams and bigrams extracted, we reduced the feature dimension by extracting 100 topics from unigrams and 100 topics from bigrams using LDA topic modelling.[2]

3. Feature Visualization (fake/non-fake reviews)



MODEL SELECTION

- **Logistic Regression:** Minimize logistic loss. Have the property of being "well-calibrated".
- **Linear Discriminant Analysis:** Generative model; The label is generated from a Bernoulli distribution. Features are normally distributed given the label.
- **Multinomial Naïve Bayes:** Generative model; The label is generated from a Bernoulli distribution. Given the label, the frequency of each word in the dictionary is generated from a multinomial distribution (n = length of text).
- **Support Vector Machine:** Large-margin classifier. Minimize hinge loss.
- **Neural Network:** 3 hidden layers with ReLU activation; sigmoid activation for the output layer

- **We used 5-fold cross validation for all our models.**

RESULTS AND MODEL COMPARISON

Models	Accuracy		Area Under Curve		F1 Score	
	Training	Test	Training	Test	Training	Test
LogicReg	0.7753	0.7618	0.8304	0.7753	0.7727	0.7616
LDA	0.7771	0.7649	0.8371	0.7771	0.7749	0.7666
MNB	0.7385	0.7241	0.7969	0.7395	0.7237	0.7103
SVM	0.7127	0.6924	0.7728	0.7170	0.6838	0.6544
NN (best)	0.8244	**0.8192**	0.8891	**0.8249**	0.8250	**0.8142**

ABLATIVE ANALYSIS

Features	Accuracy	AUC	F1 score
Reviewer + Structural + Unigram + Bigram	0.8192	0.8249	0.8142
Structural + Unigram + Bigram	0.8166	0.8195	0.8137
Unigram + Bigram	0.8011	0.8201	0.8037

With unigrams and bigrams as our baseline features, neither structural features nor reviewer-centric features contributes much to the model improvement.

DISCUSSION AND FUTURE WORK

Contrary to our expectation, adding reviewer-centric features does not boost the model performance by a significant amount. We believe it is because most reviewers only have a single review record, which makes most of our reviewer-centric features poorly defined (e.g. maximum number of reviews in a day, standard deviation of ratings). A larger dataset with more reviews attributed to each reviewer would make these reviewer-centric features more powerful. Moreover, many other reviewer-centric feature, including geo-locations and IP addresses, can also be incorporated in our model.

REFERENCES

[1].Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y Zhao. Automated crowdturfing attacks and defenses in online review systems.

[2].David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003