Discrepancy detection in Yelp reviews

Shuo Ding, Zhanhao Liu, Yanan Wang, Linhan Wei

Content

- Background
- Motivation & Objective
- Proposed Work
- Progress
- Future Work

Background

- Information impacts decision
    - Massive information is generated every day
    - People make decisions based on what information they obtain

- Fake reviews becomes everywhere
    - Study shows that nearly **16%** of all Yelp restaurant reviews in the metropolitan Boston area are fake. [1]

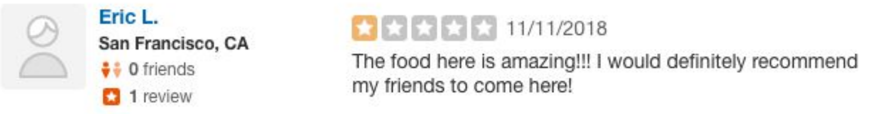
[1]: Luca, Michael, and Georgios Zervas. "Fake it till you make it: Reputation, competition, and Yelp review fraud." Management Science 62.12 (2016): 3412-3427.

Motivation & O

- Two types of dis
  - The carelessne
  - The deliberate

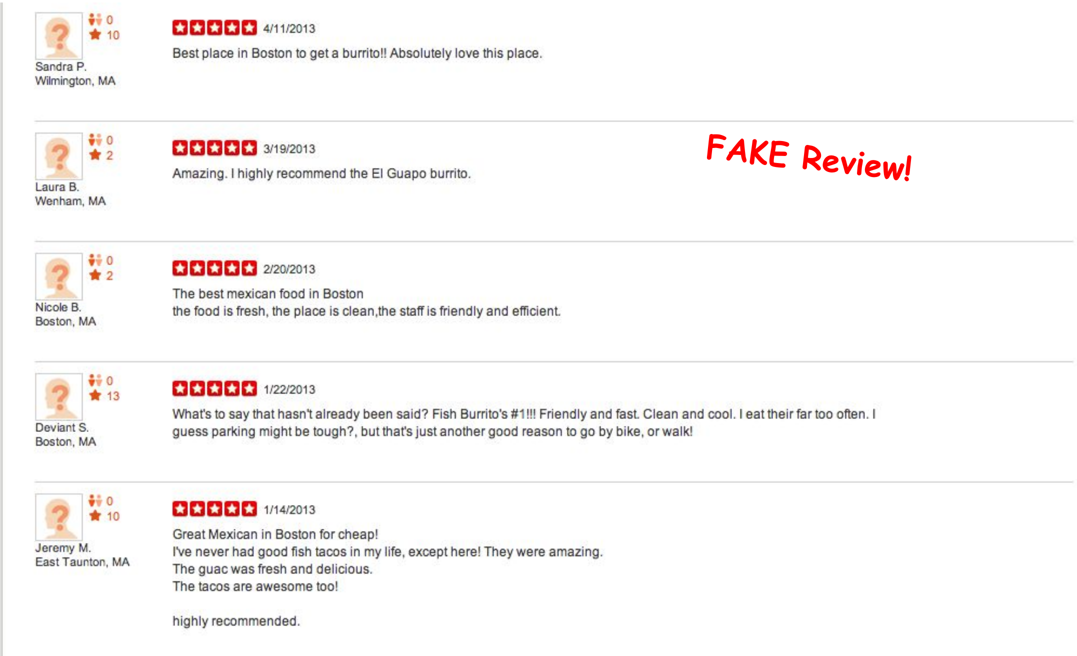
- Detect invalid or ratings based on

Discrepancy due to user carelessness



Eric L.
San Francisco, CA
0 friends
1 review

★☆☆☆☆ 11/11/2018
The food here is amazing!!! I would definitely recommend my friends to come here!



FAKE Review!

Sandra P.
Wilmington, MA
0
10

★★★★★ 4/11/2013
Best place in Boston to get a burrito!! Absolutely love this place.

Laura B.
Wenham, MA
0
2

★★★★★ 3/19/2013
Amazing. I highly recommend the El Guapo burrito.

Nicole B.
Boston, MA
0
2

★★★★★ 2/20/2013
The best mexican food in Boston
the food is fresh, the place is clean,the staff is friendly and efficient.

Deviant S.
Boston, MA
0
13

★★★★★ 1/22/2013
What's to say that hasn't already been said? Fish Burrito's #1!!! Friendly and fast. Clean and cool. I eat their far too often. I guess parking might be tough?, but that's just another good reason to go by bike, or walk!

Jeremy M.
East Taunton, MA
0
10

★★★★★ 1/14/2013
Great Mexican in Boston for cheap!
I've never had good fish tacos in my life, except here! They were amazing.
The guac was fresh and delicious.
The tacos are awesome too!

highly recommended.

Proposed Work

- Sentiment Analysis & Rate Prediction Based on the Text Reviews
  - Build a classification model to predict sentiment and rating based on text reviews
- Automatic Discrepancy Detector for Yelp Reviews
  - Use the classification model to detect fake reviews
- Machine-generated Review Detector
  - Train a new classification model to tell if a review is truthful or generated by machine
- Human-written Fake Review Detector
  - Train a new classification model to tell if a review is truthful or written by people deliberately

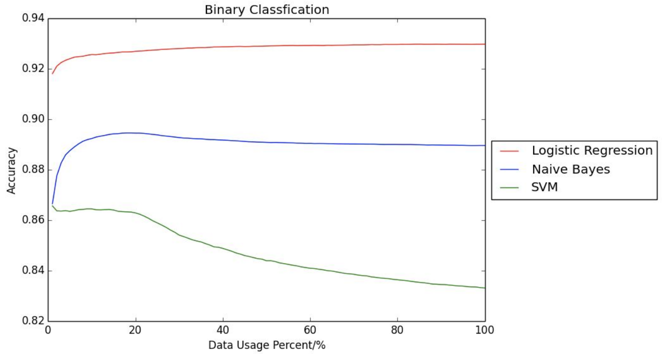
Sentiment Analysis/Rate Prediction based on the text reviews

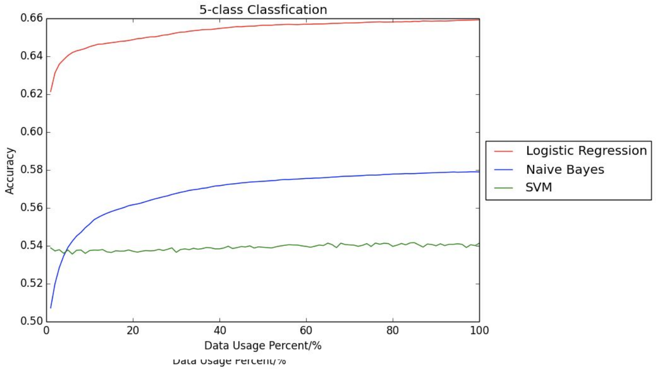
- Yelp Dataset: 6,000,000 text reviews with rating from 1 to 5
- Preprocess the Data
    - Using ***CountVectorizer*** from ***sklearn.feature_extraction.text,*** review → word tokens
- Extract Feature Vectors
    - Using ***the bag of words representation,*** word → unique ID
- Build TF-IDF Transformer
    - Using the transformer to calculate the weight of each word by using the ***tf-idf*** statistic
- Make Classification
    - Making binary classification & 5-class classification separately by using ***Logistic Regression***, ***Naive Bayes*** and ***SVM***.

Sentiment Analysis/Rate Prediction based on the text reviews

- Sentiment Analysis
  - Binary Classification (Positive & Negative)
  - Logistic Regression, Naive Bayes, SVM

- Rate Prediction
  - 5-class Classification (1, 2, 3, 4, 5)
  - Logistic Regression, Naive Bayes, SVM

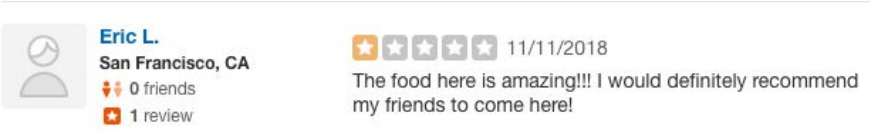

Binary Classfication


5-class Classfication

Automatic Discrepancy Detector

Motivation: The validity of the reviews

Discrepancy due to user carelessness



Eric L.
San Francisco, CA
0 friends
1 review

★☆☆☆☆ 11/11/2018

The food here is amazing!!! I would definitely recommend my friends to come here!

Automatic Discrepancy Detector

- Methodology
- Evaluation setting
- Performance Metrics
- Experimental results

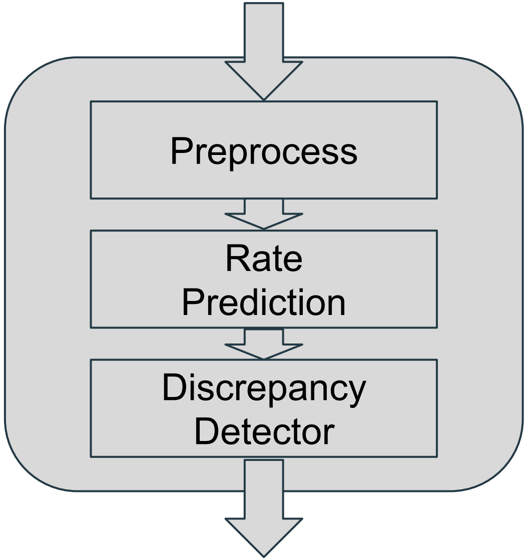
Automatic Discrepancy Detector

- Methodology

  Using the prediction model in part 1

  Designing a metric to measure

  How confidence to say the review is valid



Automatic Discrepancy Detector

- Performance Metrics

  Predict label?

  Probability?

  Distance?   Actual rating & predict rating

Automatic Discrepancy Detector

- Evaluation setting

Dataset: 6 million reviews

80% → training prediction model

20% + 10000 discrepant data (1%) → testing detector

Sorting and find 10000 reviews with the highest error metrics

Generating discrepant review:

Choose 10000 reviews in testing data.

Modify their label X

with a bias i, which is uniformly chosen from [0, 1, 2, 3]

X -> (X + i) mod 5 + 1

Automatic Discrepancy Detector

Straightforward -- Label

for review with rating X

predict(review) = i

Using $|X - i|$ as metrics

Bad idea!

$X - i \in [0, 1, 2, 3, 4]$

Automatic Discrepancy Detector

With probability

for review with rating X

predict(review) = [p1, p2, p3, p4, p5]

Using Px as metrics

7099/10000

Accuracy = 70.99%

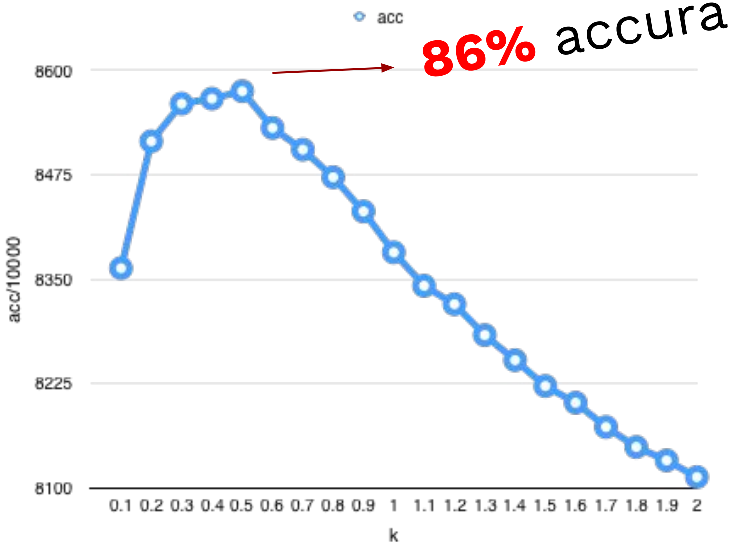
Automatic Discrepancy Detector

With probability and distance

for review with rating X

predict(review) = [p1, p2, p3, p4, p5]

Using $\Sigma pi|X - i|^k$ as metrics



86% accuracy!

Automatic Discrepancy Detector

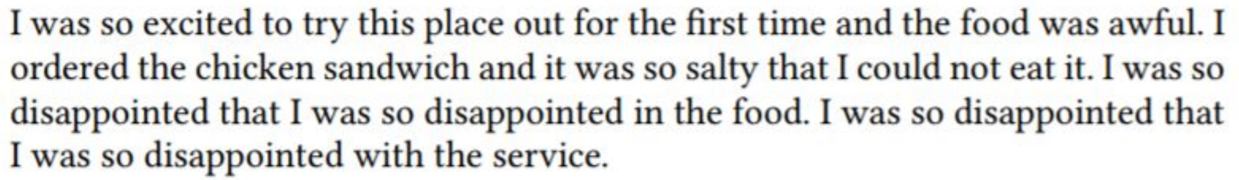
Can we do better?

Hard case

Bias is small, e.g. $1 \longrightarrow 2$

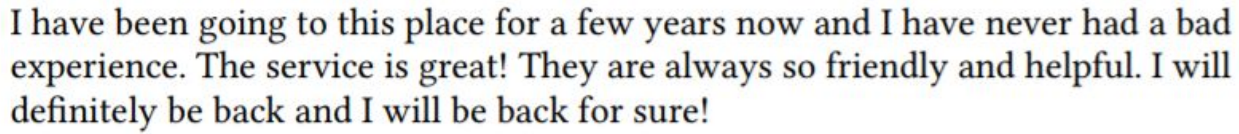
Machine-generated Review Detector

A one-star review:



I was so excited to try this place out for the first time and the food was awful. I ordered the chicken sandwich and it was so salty that I could not eat it. I was so disappointed that I was so disappointed in the food. I was so disappointed that I was so disappointed with the service.

A five-star review:



I have been going to this place for a few years now and I have never had a bad experience. The service is great! They are always so friendly and helpful. I will definitely be back and I will be back for sure!

- Yelp Restaurant Reviews Generator [1]
  - Use Recurrent Neural Network model to create human quality restaurant reviews

[1]: Yao, Yuanshun, et al. "Automated crowdturfing attacks and defenses in online review systems." Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017.

Machine-generated Review Detector

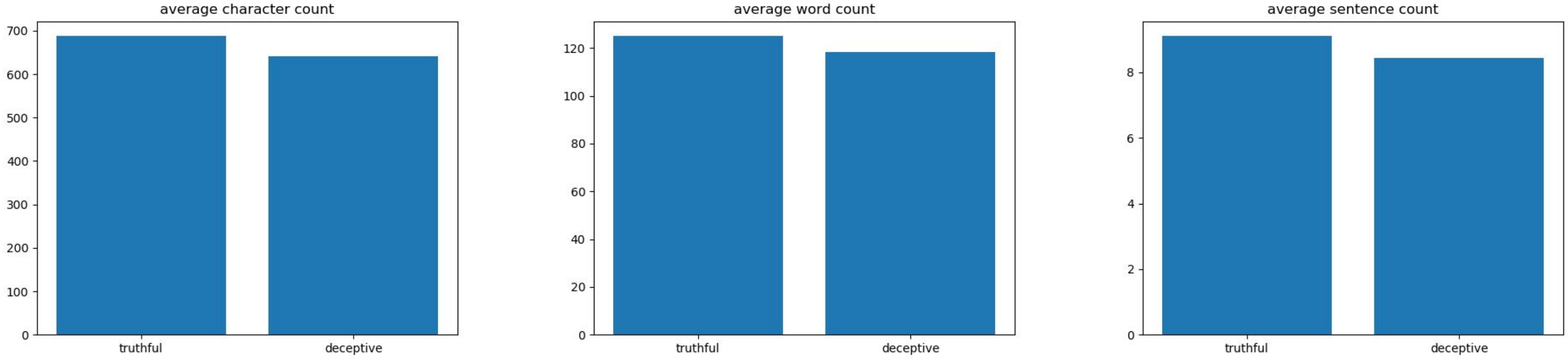
- The goal
  - Our goal is to build a review detector which can detect machine-generated reviews.
  - Our model should be general enough to handle different generators.
    - Use several different generators to train the same detector
- Some observations
  - Generally, we do not need to consider rating numbers in this task.
  - To obtain better results, we might need to add more features.

Feature Analysis

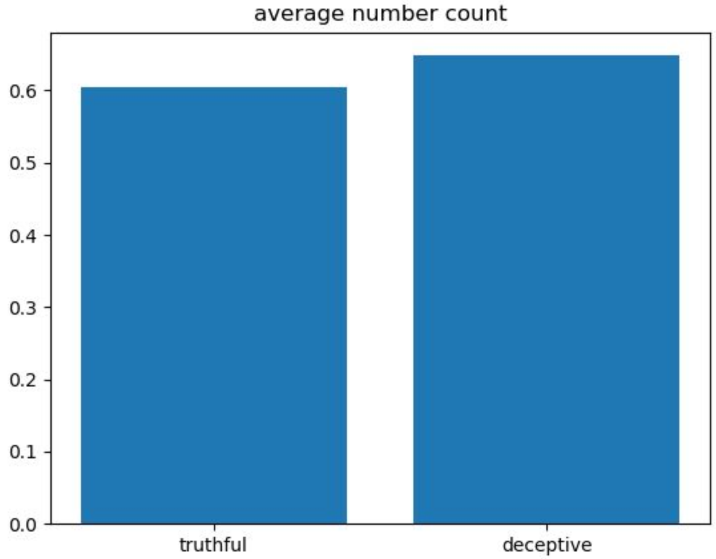
- Methodology
  - Review Features [1]
    - Character Count
    - Word Count
    - Sentence Count
    - Number Count
    - Unique Word Percentage

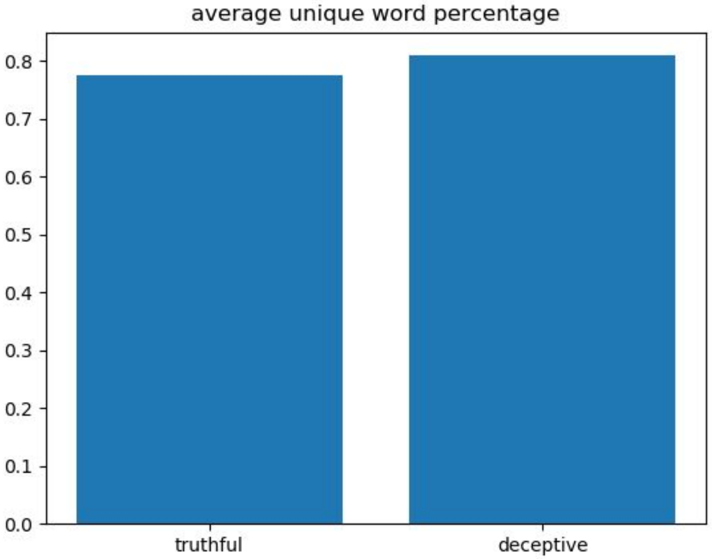
[1]: Wang, Zehui et al. “Fake Review Detection on Yelp.” (2017).

Machine Generated Review Features

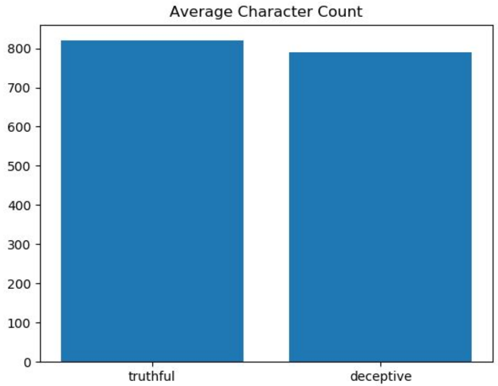


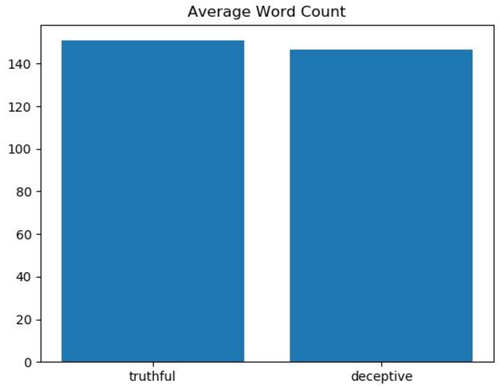
Machine Generated Review Features

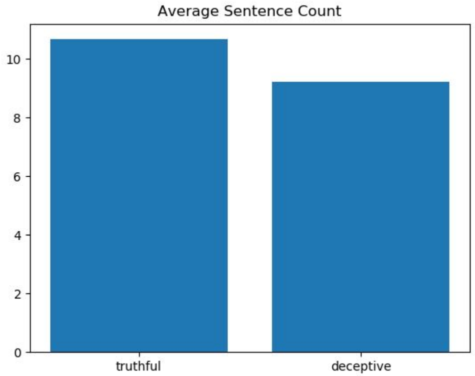




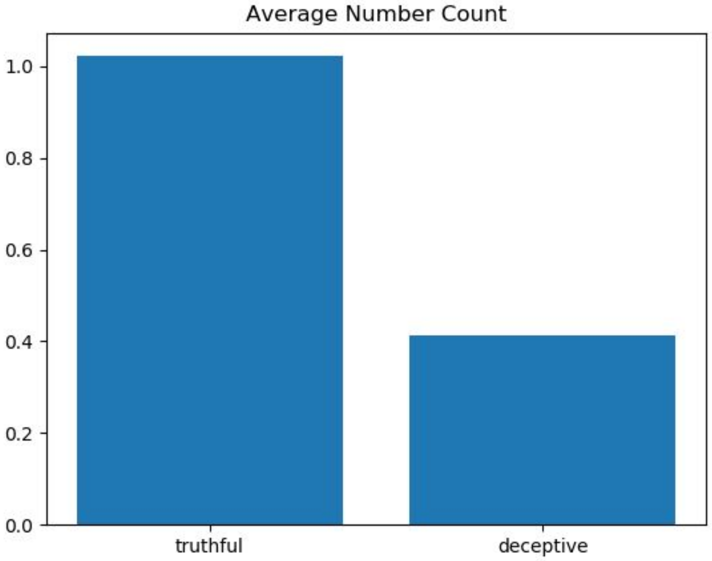
Human-written Fake Review Features

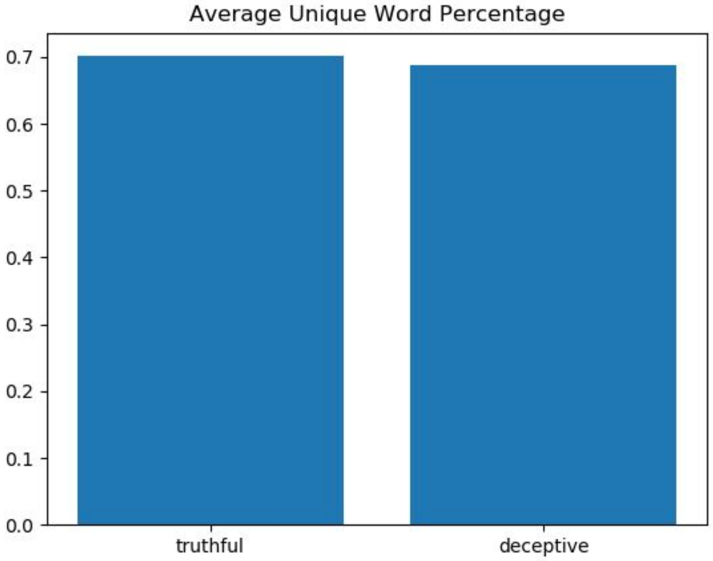

Average Character Count


Average Word Count


Average Sentence Count

Human-written Fake Review Features


Average Number Count


Average Unique Word Percentage

Machine Generated Review

Dataset

1300 truly reviews

780 generated reviews

Fake Review Machine-Generator

- Character-based Recurrent Neural Network
  - Give the RNN a huge chunk of text
  - Ask it to model the probability distribution of the next character in the sequence given a sequence of previous characters

[2]: http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Machine Generated Review

Model

Based on the model we use in the task 1

SVM, LR, MLP, RF

Adding several new features we explored

Machine Generated Review Results

SVM: accuracy = 0.709

	Precision	Recall	F1 Score
Deceptive	0.769	0.317	0.449
Truthful	0.698	0.943	0.802
avg / total	0.734	0.630	0.626

LR: accuracy = 0.698

	Precision	Recall	F1 Score
Deceptive	0.762	0.280	0.409
Truthful	0.687	0.948	0.797
avg / total	0.725	0.614	0.603

*All experiments are conducted under a 5-folds cross validation

Machine Generated Review Results

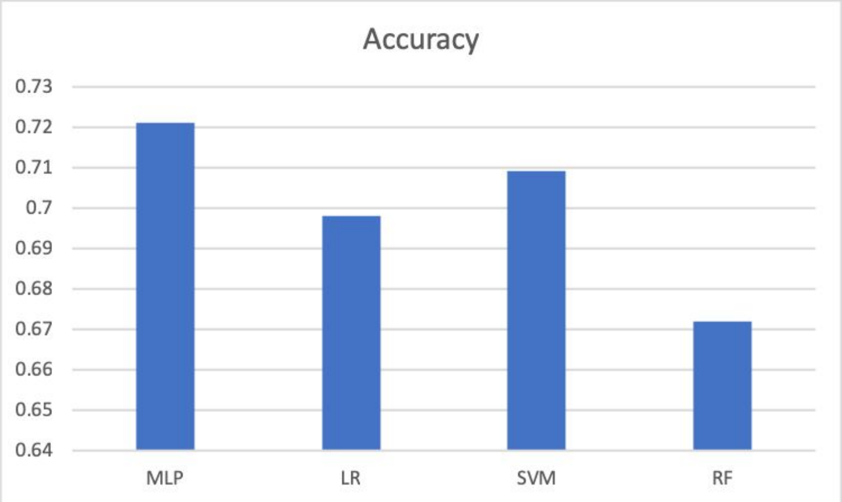
MLP: accuracy = **0.721**

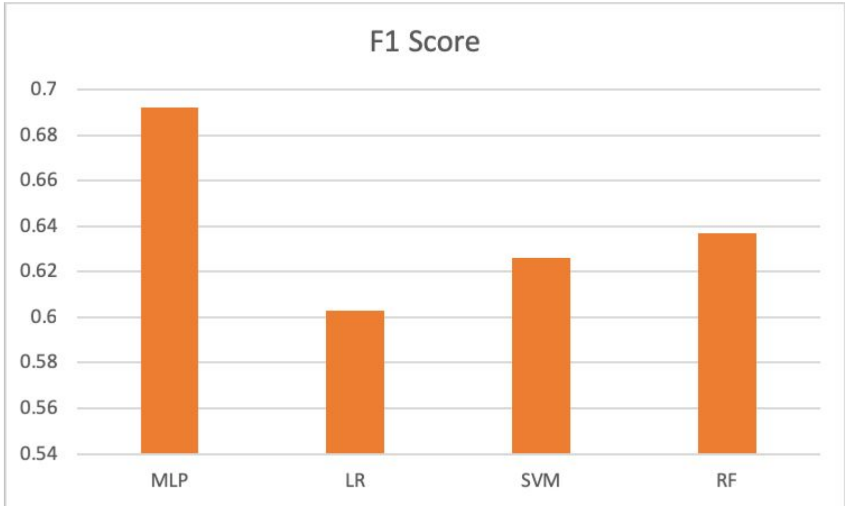
	Precision	Recall	F1 Score
Deceptive	0.650	**0.552**	0.597
Truthful	0.754	0.822	0.786
avg / total	0.702	0.687	**0.692**

RF: accuracy = 0.672

	Precision	Recall	F1 Score
Deceptive	0.574	0.483	0.524
Truthful	0.717	0.786	0.750
avg / total	0.646	0.634	0.637

Machine Generated Review Results





Possible Causes of low accuracy

- Our model is based on the words instead of the whole sentences.
- The RNN model generates high-quality reviews
- The exploring features have little difference

Analysis of Machine-Generated Fake Review Results

- A machine-generated fake review that all classifiers succeed:
  I've been here several times. The beer selection is awesome! I'm glad I was dressed with the long manicure nearby. I give them another shot, the next night had the specials a lot. It was indicative....I'll stick with earlier when we finally absolutely couldn't leave. The interior of the joint is pretty big but in the Harrah's feel of the room is almost walking to a special chair which is good, just just the best business posts in the past, but you will find Pizza Mustard Smoke Buffet. But it's a good place for a group of 15 people with a delicious communication to do a butt. We finally had a lovely kid, come vacation first rental, acceptable, trying to deal with their first time, having the lots of standards and impressions.
- Possible Causes
  - The content contains nearly no related words related to eating or hotel. Almost the whole review seems very unrelated.

Analysis of Machine-Generated Fake Review Results

- A fake review that all classifiers fail:
  This place keeps doing Italian Concept. Called a Margarita Chocolate Selection table! Highly recommend it places.
  Always mentioned it a great price for sure!
- Possible causes
  - The length of the review is too short to support the word analysis of the classifiers.
  - There are no misspelled words in this review.
  - An existed problem in this review is the syntactic error, just like sentences marked orange, which cannot be handled by the classifiers.

Analysis of Machine-Generated Fake Review Results

- A machine-generated fake review that MLP succeeds and the other fail:
  Restaurant was fresh! The service was great and the drinks were a little crispy at all and we were also disappointed.   If you are a coffee and I always do great guests but it experiences 45 years ago. It's so funny, but for great happy hours in the area, that seriously is the main staff for trying to MV. Lastly their cover was great and maybe I wouldn't be in a chain and within didn't seem to get reengo with their service. I paid $30/much and giving my bad chance of stopping after doathers. You can also bring my cell phone. Guess that's the stuff.
- Possible causes
  - Some sentences of the review have nothing to do with eating but it does contain some words related to eating. In this case, MLP shows the power dealing with such complex problem.

Human-written Fake Review

- Dataset
  - Deceptive Opinion Spam Corpus: A corpus of truthful and deceptive hotel reviews
  - 400 truthful positive reviews from TripAdvisor [1]
  - 400 deceptive positive reviews from Mechanical Turk [1]
  - 400 truthful negative reviews from Expedia, Hotels.com, Priceline, TripAdvisor, Yelp [2]
  - 400 deceptive negative reviews from Mechanical Turk [2]

[1] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. 2011. 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.
[2] M. Ott, C. Cardie, and J.T. Hancock. 2013. Negative Deceptive Opinion Spam. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Human-written Fake Review Results

NB: accuracy = 0.857

	Precision	Recall	F1 Score
Deceptive	0.812	**0.929**	0.866
Truthful	0.917	0.785	0.846
avg / total	0.864	0.857	0.856

LR: accuracy = 0.878

	Precision	Recall	F1 Score
Deceptive	0.871	0.886	0.879
Truthful	0.884	0.869	0.876
avg / total	0.878	0.877	0.877

*All experiments are conducted under a 5-folds cross validation

Human-written Fake Review Results

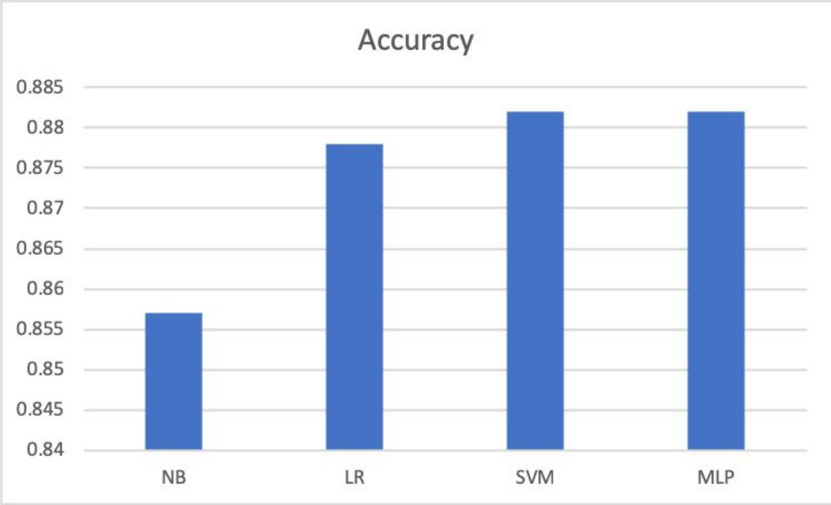
SVM: accuracy = **0.882**

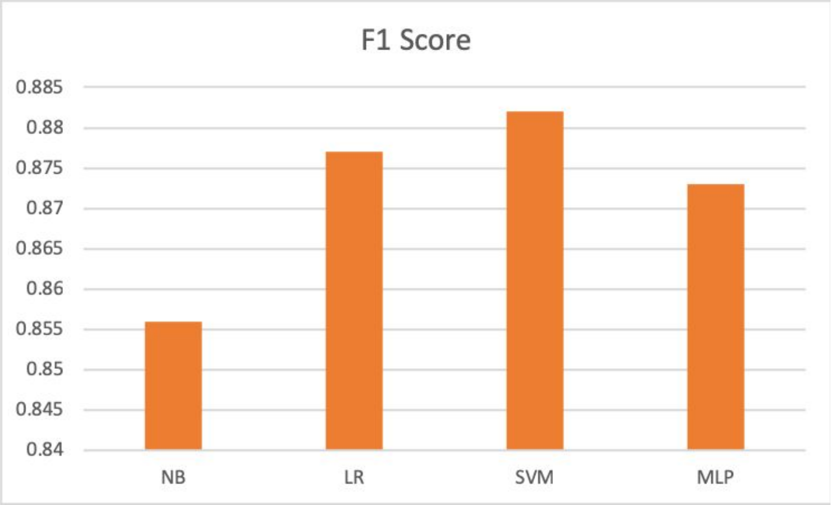
	Precision	Recall	F1 Score
Deceptive	0.889	0.874	0.881
Truthful	0.876	0.891	0.884
avg / total	0.883	0.883	**0.882**

MLP: accuracy = 0.873

	Precision	Recall	F1 Score
Deceptive	0.864	0.885	0.875
Truthful	0.882	0.861	0.872
avg / total	0.873	0.873	0.873

Human-written Fake Review Results





Analysis of Human-written Fake Review Results

- A human-written fake review that all classifiers succeed:
  Hard Rock Hotel boats that they have the best of the high scale accommodations for your business or pleasure stays. What they fail to disclose is just how expesive your stay will be. The service tends to be heavily influenced by the wealth of the patrons, service is more redily available and friendly the more you make.The rooms are furnished with attractive but old furnishings. The laundry service leaves more to be desired. The sheets have a gray tint to them and the rooms smell of tobacco. I recently stayed at the Hard Rock Hotel and was diappointed with not only the service but the cleanliness of the room from the bathroom to the beding.Your stay at any hotel should be memorable and restful-after all isnt that what we pay for?
- Possible Causes
  - The length of the review is long enough to support the word analysis of the classifiers.
  - There are several misspelled words marked red in this review.

Analysis of Human-written Fake Review Results

- A fake review that all classifiers fail:
  Hotel is located 1/2 mile from the train station which is quite hike when you're traveling with luggage and/or kids. They seem to cash in on guests who arrive in private car by charging exorbitant parking/valet fees. Rooms feature either double or king sized beds; no queen beds at all. If you want a little extra leg room in your bed, the price jump from double- to king-sized is stiff. Rooms with any kind of view pay a healthy surcharge, too.
- Possible causes
  - The length of the review is too short to support the word analysis of the classifiers.
  - There are no misspelled words in this review.
  - An existed problem in this review is the syntactic error, just like sentences marked orange, which cannot be handled by the classifiers.

Analysis of Human-written Fake Review Results

- A human-written fake review that MLP succeeds and the other fail:
  This hotel was very poor with customer service. They were so worried about keeping everything up to date and perfect looking in the hotel they rarely worried about their customers, and their guests. I had to ask for my room to be serviced for cleaning, and it took a long time to get a response. Very disappointed.
- Possible causes
  - Neural Network is more power than other three classifiers.
  - This review is relatively short which makes other model fail.
  - There are plenty use of words within same word class. Since NN have several layer to deal with the input data, and our model is based on the frequency of the words, NN will be more powerful when dealing with this case

Conclusion

- Fully completed the four tasks: **Sentiment Analysis & Rate Prediction**, **Automatic Discrepancy Detector**, **Machine-generated Fake Review Detector** and **Human-written Fake Review Detector**.
  - Sentiment & Rating prediction achieve high accuracy based on the text reviews.
  - Explored and designed our own metrics to tell if the reviews are discrepancy.
  - Compared and analyzed the difference among the truthful reviews, machine-generated fake reviews and human-written fake reviews.
  - Based on the classification results, we analyzed the possible reasons behind the successes or failures.
- Future work
  - Add syntactic analysis into our model to make it more powerful.