SPOT-IT

Flag deceptive reviews

DS 501 case study -4 Team#6- Yousef Fadila, Rahul Ghadge & Praneeth Nooli

Introduction

With the invention of the web, reviews and opinions about everything are available at our fingertips. Before buying any product or availing any services we always check for its reviews and opinion of others, probably from experts or customers who have used the it. This means, anything and everything that is available on the web has a role in forming our own opinion about a particular product or service. It influences our opinion to a certain extent. Sometimes even our decision to buy any product is influenced by such reviews available on websites. We look for a product having most positive reviews but how do we know the reviews are truthful? In today's competitive world, many companies pay people to submit positive reviews and rating for their products which helps them to increase their sale indirectly. But from customer's point of view this is misleading. Once we went to a restaurant after reading good reviews online, about its food ambience and service, but to our dismay none of this was up to the mark and way below our expectation. At that time, we realized we are being tricked and like this any customer can be deceived. Because, human's brains suffer from a "truth bias" i.e. assuming what they are reading is true until they find evidence to the contrary. Indeed, this is the motivation of our work in this case-study. To predict and classify truthful or genuine reviews available online from the ones which are deceptive or fake.

In this report, we will detail the process and goals of detecting fake positive reviews using three primary objectives.

Objective 1: The Business Part. Precisely describe the business problem fake review detection aims to solve, discuss why this problem is important, and explore how our data-science techniques could drive business value.

Objective 2: The Math Part. Formulate the business problem as a math problem, and develop a mathematical solution for implementation.

Objective 3: The Hacking Part. Develop a prototype of fake positive review detector. Develop general solution that works for any website based on the text analysis itself plus refine this solution to increase accuracy for all major website. In the context of this project we will refining the solution using data from yelp challenge dataset.

Objective 1: The Business Part

The Business problem to solve

The Web has greatly enhanced the way people perform certain activities such as shopping etc. in order to find information, and interact with others. Today many people read/write reviews on merchant sites, blogs, forums, and social media before/after they purchase products or services. Examples include business reviews on Yelp, product reviews on Amazon, hotel reviews on TripAdvisor, and many others. Such user-generated content contains rich information about user experiences and opinions, which allow future potential customers to make better decisions about spending their money, and also help merchants improve their products, services, and marketing.

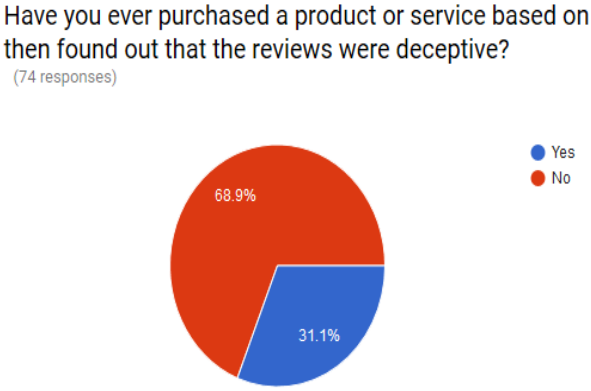
User-generated online reviews can play a significant role in the success of retail products, hospitals, restaurants, etc. However, review systems are often targeted by opinion spammers who seek to distort the perceived quality of a product by creating fraudulent reviews. In recent years, fake review detection has attracted significant attention from both businesses and the research community. We propose a fast and effective framework known as "Spot.It" in order to detect the fake positive reviews.

Why this problem is important to solve

This is problem is major concern for both customer and enterprises, which has been reported in many news articles. Below are some examples.

Fake Reviews in the news!

- Samsung probed in Taiwan over 'fake web reviews',
- Woman Paid To Post Five-Star Google Feedback,
- Are You Buying Reviews For Google Places?,
- The Best Book Reviews Money Can Buy,
- For $2 a Star, an Online Retailer Gets 5-Star Product Reviews,
- Amazon Glitch Unmasks War Of Reviewers
- Charges Settled Over Fake Reviews on iTunes
- Company Settles Case of Reviews It Faked
- TripAdvisor warns consumers about fake reviews
- Belkin's Development Rep is Hiring People to Write Fake Positive Amazon Reviews
- A Fake Amazon Reviewer Confesses



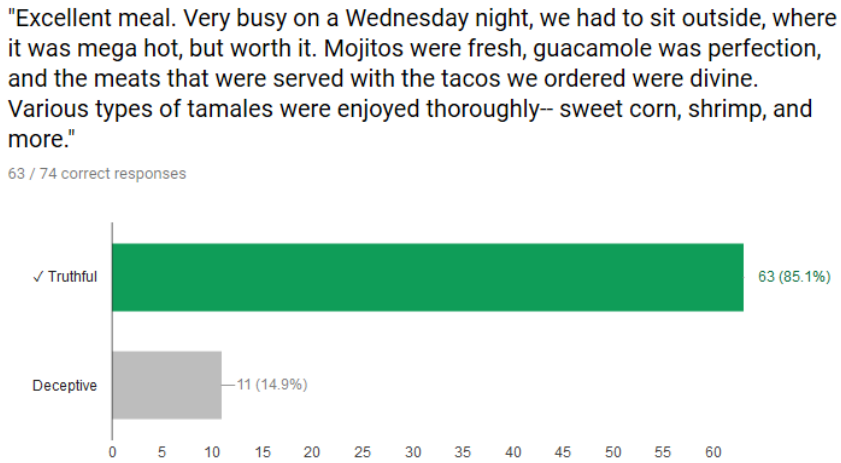
According to a report of a survey sent to WPI CS/DS graduate students, over 30% of the participant admit that they were victim of deceptive reviews.

Online reviews are increasingly used by individuals and organizations to make purchase and business decisions. Positive reviews can render significant financial gains and fame for businesses and individuals. Unfortunately, this gives strong incentives for imposters to game the system by posting fake reviews to promote or to discredit some target products or businesses. In the past few years, the problem of spam or fake reviews has become widespread, and many high-profile cases have been reported in the news. Consumer sites have even put together many clues for people to manually spot fake reviews. There have also been media investigations where fake reviewers blatantly admit to have been paid to write fake reviews. The analysis on fake reviews states that many businesses have tuned into paying positive reviews with cash, coupons, and promotions to increase sales. In fact the menace created by rampant posting of fake reviews have soared to such serious levels that Yelp.com has launched a “sting” operation to publicly shame businesses who buy fake reviews. For reviews to reflect genuine user experiences and opinions, detecting fake reviews is an important problem.

In addition, over 85% percent of people show interest in automatic deceptive detector solution that would mark or filter out deceptive reviews from web pages.

The ideas behind Fake review detection to solve this problem

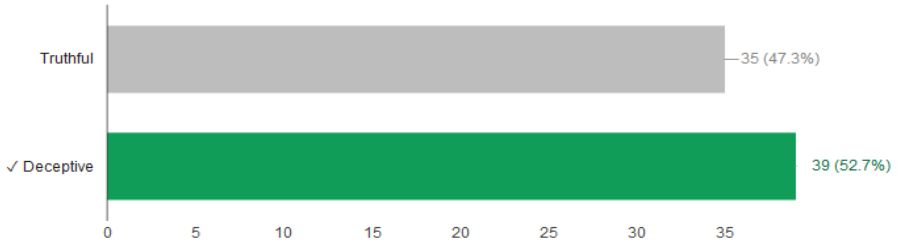
“A research article by Cornell university states that human suffers from a “truth bias”, assuming that what they are reading is true until they find evidence to the contrary. When people are trained at detecting deception they become overly skeptical and report deception too often, still scoring at chance levels. Truth-tellers and deceivers differ in the use of keywords referring to human behavior and personal life, and sometimes in features like the amount of punctuation or frequency of "large words." In parallel with the analysis of imaginative vs. informative writing, deceivers use more verbs and truth-tellers use more nouns."



We have tested this claim by surveying WPI, CS/DS graduate student. Survey includes true reviews taken from high-rating business and trusted user, and fake, deceptive reviews written by turkers from www.mturk.com.

My friend and I visited Joe's after hearing much praise about the restaurant. Upon entering the restaurant, we were immediately greeted upon and it did not take long for us to be seated and waited on. We started with calamari (a must-have!) and oysters. The seasoning on the calamari was just right and the sauce was a great complement. Then we ordered the crabs, which far exceeded my expectations. They were a good portion and were fresh and broiled to perfection. And of course, I can't forget dessert! I have such a sweet tooth, so I know a good dessert when I come across one, and I must say, their blueberry pie was absolutely delicious. It was the perfect way to end the perfect meal.

39 / 74 correct responses

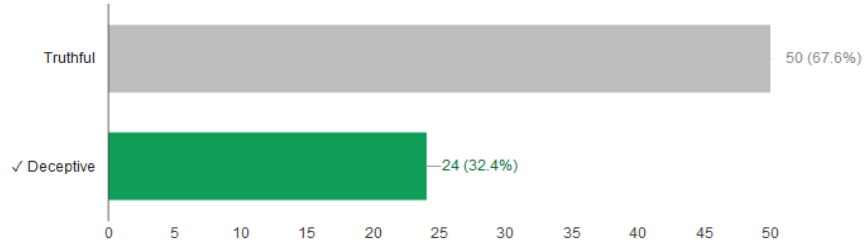


But when it comes to deceptive reviews, the result was surprising! For example 47.3% of the participant classify this fake-deceptive review as real one.

In other question, the result was even worse.

"My husband and I recently stayed at your Hard Rock Hotel in Chicago, and what a fantastic experience it was. We stayed in one of the 'Extreme Suites' and let me tell you it was beautiful. The furniture and paintings were amazing and the view was to die for. The staff treated us really well and met all of our needs and then some. Thanks so much!"

24 / 74 correct responses



Because, 67.6% of the participants thought that this fake-deceptive review is genuine.

This opens the door to develop a general, deceptive ranking solution to detect deceptive reviews based on the text analysis itself regardless of the website using the features.

So, based on the above mentioned idea we decided to build a general solution to detect deceiving reviews based on the test itself and the differentiation between imaginative vs informative writing. The solution can be used as its own or as a stage in more specific solution like targeting specific website.

In our implementation, as 1st phase we will build 2 models.

1) General model that gives real/fake categorical identifier based on content and features extracted from the content. As 1st phase, the model will be trained using restaurant reviews data only. In addition, the model will be deployed and exposed to public using REST API.

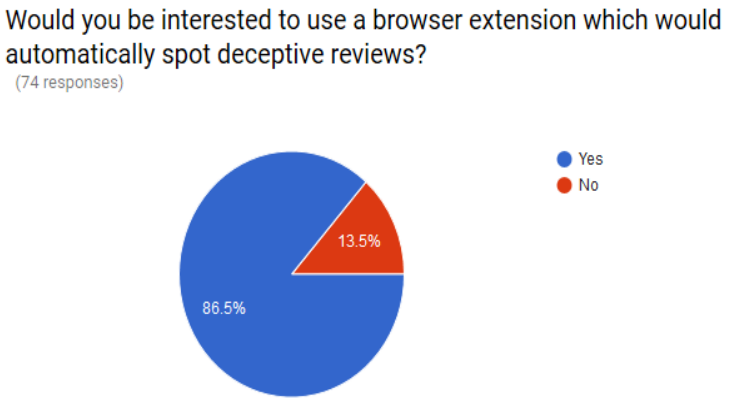
2) Specific model, This model uses anomaly detection method to develop a model to detect the "outlier" reviews based on user/business metadata as first step, then use the general model to refine the solution and reduce false positive errors. I.e. fake positive reviews.

Differences that can be made via our data science approach

Several research papers and the survey we conducted pose a very interesting question- "Are computers better than Humans in detecting deceptive reviews?" and results prove the claim that humans suffer from truth bias. Even though we deal with natural language problem of classifying short reviews, which in theory, trivial for humans but difficult to computers, The survey shows that well training data based solution would beat human ability to detect deceptive text. The mentioned research suggested that computers would be able to reach high accuracy in text based solution.

In addition to the general-text-based solution. The specific anomaly detection solution introduces analyzing a lot of parameters and metadata that can't be done without data-science techniques.

This idea deserves investment from the Sharks



Following the survey result, 86.5% show interest in a solution that could automatically spot deceptive reviews from them. This clearly shows the public concern from being scammed by deceptive-positive reviews.

Based on what we see that this product would drive value for large number of population, in addition to the end-user target, like - businesses that share user's capacity to other users (Peer-to-Peer). E.g. - Airbnb, Turo are based on peer-to-peer reviews and trust, a solution to automatically detect deceptive reviews could be a core of interest to this rapid-growing market.

In addition, the same technique used in our product could be tuned to different sectors with high growth potential. such as detecting fake news, rumors in social media.

Objective 2: The Math Part

Math behind the general solution, text based approach:

Features from the text classifiers, ngram(1,3), plus features which differentiates informative writing and imaginative writing, (ratio of verbs, nouns, punctuations) are used to train support vector machine classifier which have performed well in document classification problems.

$$\hat{y} = sign(\vec{w} \cdot \vec{x} + b)$$

For simplicity, we train the system using linear-SVM which learn a weight vector W and bias term b such that a review X can be classified.

Math and motivations behind the specific 2 steps solution, anomaly detection followed by text based classification method:

1) Not a simple outlier detector. Basing the solution to detect outliers on the metadata only would result in having high rate of false positive, these outliers could be legal users/ Businesses. For example, new user/business, user has small number of reviews, very popular businesses that have high-rate of positive reviews and so.

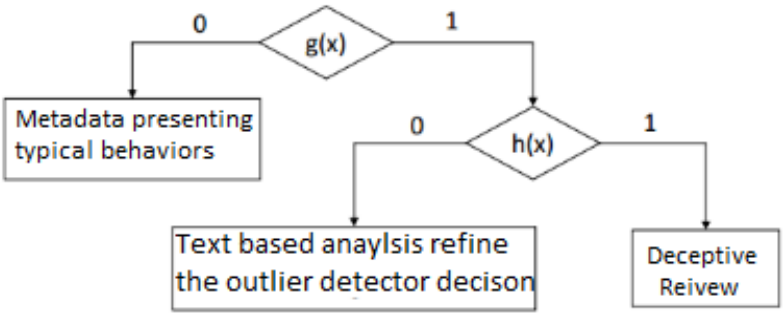
2) Not a text based analysis only . Having base of the solution on text analysis only would neglect many features that are very helpful to detect deceptive reviews. Such as the average rate of reviews of the specific user, number of online friends/followers the user have and so on. The research proves that text based analysis could reach high rate of 90% accuracy, but the problem is that the mechanical turks can be also get advanced as the text based filter get advanced., The current situation is that most deceptive reviews can be detected using the differentiation between informative vs imaginative. But if the solution is only based on this fact, the mechanical turk can easily bypass by writing reviews using informative style, less large word or maybe be copying legal, real reviews that the algorithm detect as real and use them with small tuning to review other businesses. The anomaly detection step would prevent such problem.

Below is the mathematical representation of the solution.

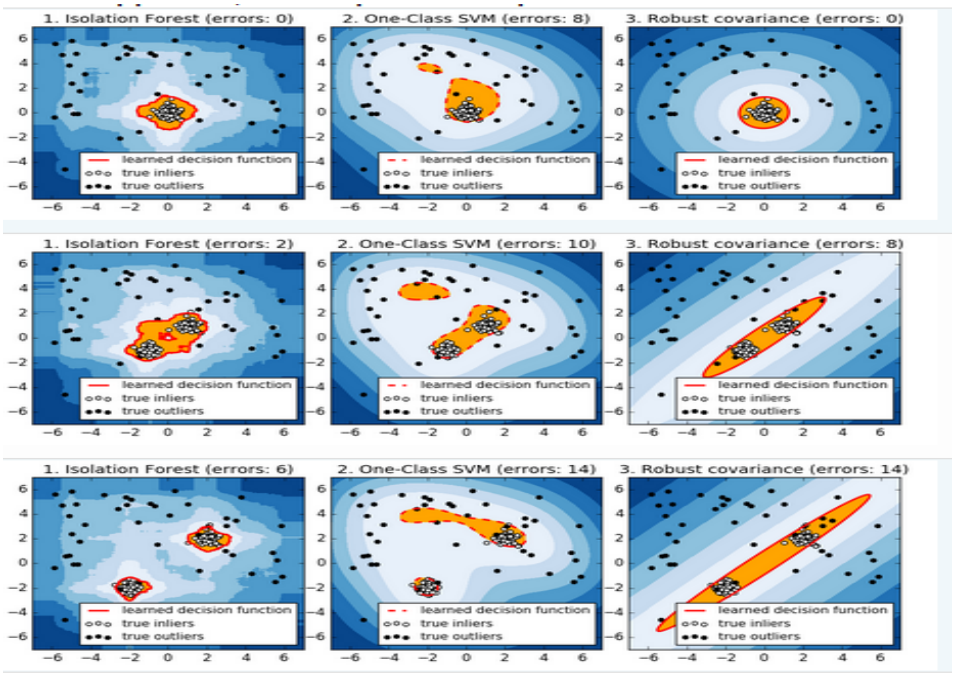
$$f(x) = \begin{cases} 0 & \text{if } g(x) = 0 \\ h(x) & \text{if } g(x) = 1 \end{cases}$$

g(x) → anomaly detection function

h(x) → outlier detector - text classifier function



Flow→ So a given review will be classified as a deceptive one if it has been detected as an outlier by the anomaly detector and then classified as a deceptive review by the text classifier.



For the anomaly detection step we use the **Elliptic Envelope** algorithm. It is logical to consider the legal user behavior having similarities that could be gather them in one region[TODO: add visualization], other methods like One-class SVM which is a novelty checker and require a pure data or Isolation Forest which could result having the new users as who give single reviews as “legal forest”. While we want to detect these users as outliers, it doesn’t mean we automatically classify their reviews as deceptive. We intentionally allow some false-positives in this phase as the result will be refined using text based analysis. For that we choose the Elliptic Envelope rather than Isolation Forest and One-class SVM.

Objective 3: The Hacking Part.

Though "Spot It" has not undergone full scale software and product development, we have executed a prototype of the "Spot It" in order to validate the underlying ideas.

Data Collection: In order to build the model for Spot It, we have collected two datasets. One is the Yelp dataset which has 2 million Reviews and the other dataset is the Cornell Deceptive Review Dataset. From 2 million reviews collected from yelp, we are using reviews related to restaurant business category.

Data Analysis: As mentioned earlier, we will follow a 2-step approach which is as mentioned below for building the "Spot It" framework.

For detecting the outliers, we first built an anomaly detector based on the Yelp dataset restaurant reviews in which the following features are used for building the model.

- user_review_count
- user_average_stars
- user_Friends_count
- Business_rate
- Business_review_count

For the ease of implementing the model, we considered only positive reviews from the yelp dataset and are trying to spot fake positive reviews. So in order to extract the positive reviews, we filtered out negative and neutral reviews by considering only the reviews which have a rating of 4 stars and greater. So, by considering the above mentioned features we build the anomaly detector model with the help of Elliptic Envelope for detecting the outliers.

In order to classify a given review as a deceptive or not, we built a text classifier with the help of Linear Support Vector Classifier by training the model with Cornell Deceptive Review Dataset using the features as mentioned below-

- amount of punctuation
- total verbs - total nouns
- length of the review

	precision	recall	f1-score	support
deceptive	0.82	0.95	0.88	19
truthful	0.95	0.82	0.88	22
avg / total	0.89	0.88	0.88	41

Confusion Matrix

```
[[18  1]
 [ 4 18]]
```

The classification report and the Confusion Matrix for the test results of the text classifier are as shown on left side.

After we build the two models, we combine the two models and classify a particular Yelp review as a fake positive review only if the review was detected as an outlier from the anomaly model as well classified as a fake review from the text classifier.

Based on this idea we built two types of detectors.

1) An offline detector which detects the fake reviews based on the above mentioned 2 step approach from Yelp offline database json files
2) An online real time detector which takes the reviews from Yelp website given the business Id with the help of Yelp rest API and classifies the review based on the 2-step approach.

The following is the demonstration of online real time detector which takes a review from Yelp website given the business Id as 'shawarma-palace-worcester'

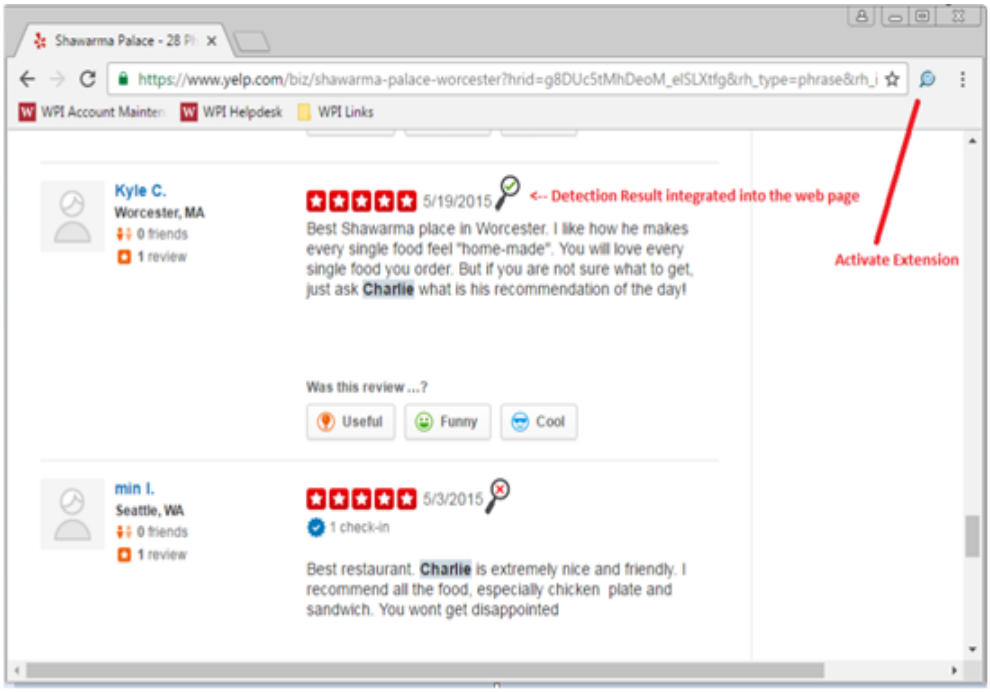
Mentioned Review - "I've been meaning to stop here since I moved to the neighborhood over two years ago. I finally had the chance yesterday and it FAR exceeded my expectations...."

Result - "Real"

Explanation - As mentioned previously the deceptive review contains more verbs and it over exaggerate the things by trying to be imaginative rather than informative. Here in the above-mentioned review we did not observe any such things which considers to be a deceptive review. So, our detector classified it as a real review.

Browser Extension and how our analysis support our business proposition:-

Our business proposition is to help customers to make better decisions by not to mislead from the reviews they read. In order to achieve this, we proposed a 2-way approach by building an anomaly detector using the Elliptic envelope which was discussed in the Math part and a Text classifier model. Our end result is to build a product which helps the user in determining the reviews they have encountered are deceptive or not. So, for that purpose we created a browser extension. Its working can be seen in the below mentioned figure.



So, the above mentioned reviews are for a restaurant named "Shawarama Palace Worcester" which are mentioned in the yelp website. We can see that the browser extension classified the second review as Deceptive.

Conclusion

In this report, we described about the problems that were experienced by the people due to the fake reviews, and then introduced to our new product, “Spot It”. Spot It makes use of data science techniques to predict whether a given review is deceptive or not. So first we built an anomaly detection model using elliptic envelope algorithm in order to determine whether the given review as an outlier or not. Once a review has been detected as an outlier, further it will be passed to a text classifier and is considered to be a deceptive review if the text classifier classifies it as a fake review. In order to check up to date reviews from the yelp website, we built an online real time detector using the yelp rest API apart from the offline detector which takes the reviews form the Yelp offline databases.