|  |
| --- |
| Springboard Capstone Two |

Deception Detective

Identifying Fake Reviews using Natural Language Processing

11/7/2017



# Contents

1. The Problem
2. The Data
3. Natural Language Processing
4. The Models
5. The Results
6. Future Enhancements
7. References

### The Problem

A recent Forbes Article suggests that the fake review problem is getting worse, not better, despite efforts to minimize the practice. A recent study by Harvard Business School estimates that as many as 20% of online reviews may be fraudulent.[3]

Consumers rely on online reviews to make decisions about which products and services to purchase. Fake opinion reviews are a big problem because they make customers doubt the accuracy of online reviews. Research has shown that machine learning algorithms are better at identifying fake opinion reviews than human judges [1].

One of these two reviews is a genuine Yelp review and one is a fake review. Can you tell which is which?

1. 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.

2. 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 definatly be back to Chicago and we will for sure be back to the James Chicago.

If you have trouble identifying the fake one, you are not alone (#2 is the fake.). In a study done by Mylie Ott [1], human judges identified fake opinion reviews correctly with about the same accuracy as a random pick.

Humans are poorly equipped to distinguish between fake and genuine reviews. This project describes a use-case where machine learning algorithms using natural language processing are better at identifying opinion spam than human judges. For this reason, machine learning algorithms which can detect fake reviews would be useful for any commercial endeavor which wants to assure that their customers have access to genuine reviews. Review sites such as Yelp and TripAdvisor could use the models to flag potential fake reviews.

### The Data

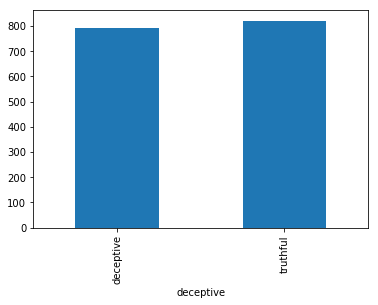
The dataset contains 1600 reviews; each review is tagged with four identifiers: 1. whether it is deceptive or truthful, 2. the name of the hotel (one of twenty of the most popular hotels in Chicago), 3. polarity (negative or positive) 4. the source of the review (Yelp, TripAdvisor, or Mechanical Turk). There are 400 genuine positive reviews, 400 truthful negative reviews, 400 deceptive positive and 400 deceptive negative reviews for each of 20 of the most popular hotels in Chicago.

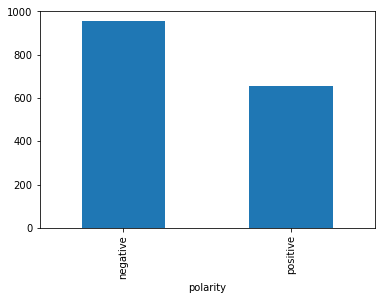
An examination of the data reveals that there are no null values, no outliers, and that the dataset is balanced between negative and positive reviews, and truthful and deceptive reviews. Each review text contains at least 150 characters or more.

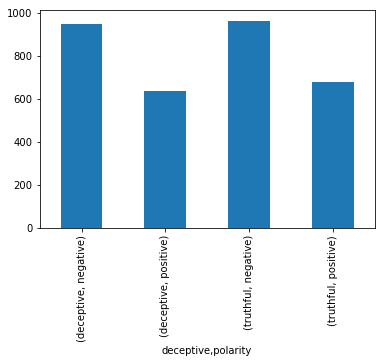
The first three rows in the dataset are shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **deceptive** | **hotel** | **polarity** | **source** | **text** |
| **0** | truthful | conrad | positive | TripAdvisor | We stayed for a one night getaway with family ... |
| **1** | truthful | hyatt | positive | TripAdvisor | Triple A rate with upgrade to view room was le... |
| **2** | truthful | hyatt | positive | TripAdvisor | This comes a little late as I'm finally catchi... |
| **3** | truthful | omni | positive | TripAdvisor | The Omni Chicago really delivers on all fronts... |

A statistical examination of the mean word length reveals that there is no statistically significant difference between the average word length of deceptive reviews and truthful reviews. The bar charts below illustrate this examination.







### Natural Language Processing

Text Frequency-Inverse Document FSrequency (TD-IDF)

First we apply a technique frequently used in data mining called TD-IDF. Each review counts as one document of our corpora, and individual word frequencies are counted for each document. The frequency vectors for each document are combined into a matrix as described below with the dimensions TxF where T is the number of documents and F is the number of individual terms in the collection of documents. The resulting matrix and is used to predict the binary class that the document belongs to, deceptive or truthful. The matrix is used as the input for Linear Support Vector Classification and Logistic Regression.

Topic Modeling

### The Models

LogisticRegression

GaussianNB

SVC

LinearSVC

RandomForestClassifier(n\_jobs = -1, n\_estimators = 500

LinearDiscriminantAnalysis

1. Logistic Regression

Logistic regression fits a logistic model to data and makes predictions about the probability of an event (between 0 and 1).

1. Naïve Bayes

Naive Bayes uses Bayes Theorem to model the conditional relationship of each attribute (the x’s) to the class variable (y, the target variable).

1. Support Vector Machines

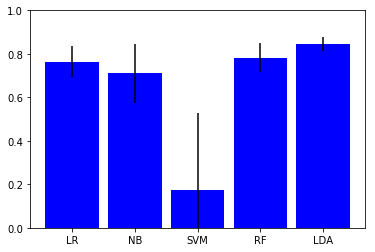
Support Vector Machines (SVM) is a method that uses points that best separate classes into two groups in order to predict classification.

1. Linear Support Vector Machine
2. Random Forest is an ensemble model which aggregates the results of a specified number of decision trees to predict the most likely classification. Decision trees are randomly generated, each prediction is summed up and the winning prediction is generated.
3. 6 Linear Discriminant Analysis The purpose of linear discriminant analysis (LDA) is to estimate the probability that a sample belongs to a specific class given the data sample itself. Applying Bayes Theorem results in. The shared covariance matrix and mean vectors are estimated from the training data.

### The Results

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | (Standard Deviation) |
| Logistic Regression: | 0.764375 | (0.072460) |
| Naïve Bayes: | 0.710000 | (0.134809) |
| Support Vector Machine: | 0.175625 | (0.351253) |
| Linear Support Vector Machine: | 0.838125 | (0.035072) |
| Random Forest: | 0.782500 | (0.067419) |
| Linear Discriminant Analysis: | 0.845000 | (0.033981) |

**Results Comparison**



## There are various ways to evaluate the success of a predictive model. Some of them are listed below:

* **Predictive Accuracy**: How many does it get right? This is generally the most important metric and is shown in the above chart. However, some other important considerations are:
* **Speed**: How long does it take for the model to deploy? Since the dataset we are evaluating is not overly large, 5 x 1600, 5 columns, or variables, and 1600 rows, or instances, the time for the model to deploy is not much of a factor. For larger datasets this would be more of an issue.
* **Scalability**: Can the model handle large datasets? Given the size of the dataset, this is not a significant evaluative issue.
* **Robustness**: How well does the model handle outliers and missing values? The dataset comes already curated within certain parameters with no outliers or missing values. Therefore robustness is not an evaluative factor.
* **Understandability**: Is the model easy to understand? Linear classification models work by dividing the data into two classes and drawing a line separating the two classes, making them some of the easiest algorithms to understand and interpret.

**For this use-case the accuracy score is the most important evaluative tool.**

And the winner is:

TF-IDF/Linear Support Vector Classification:

accuracy score: 0.89375

Runner-Up:

Topic Modeling/Linear Discriminant Analysis:

accuracy score: 0.845000

### References

[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.

[3] Woolacott 2017 <https://www.forbes.com/sites/emmawoollacott/2017/09/09/exclusive-amazons-fake-review-problem-is-now-worse-than-ever/#2fee58c47c0f>