|  |
| --- |
| 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. Analysis
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 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.

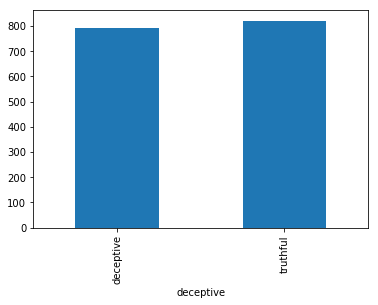
Before we proceed with our data analysis, we want to examine the data to see if it needs to be “cleaned”. Does it contain null values which will hinder our analysis, are there outliers which will skew our statistical analysis, and is our dataset unbalanced or balanced, in other words are the number of truthful reviews roughly equal to the number of deceptive reviews?

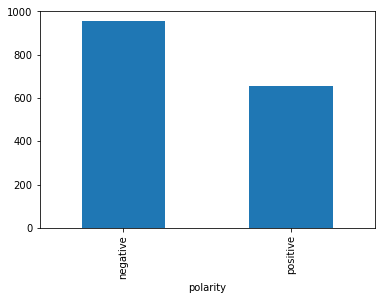
An examination of the data reveals that there are no null values, no outliers, and the dataset is balanced between negative and positive reviews, and truthful and deceptive reviews. Each review text contains 150 characters or more.

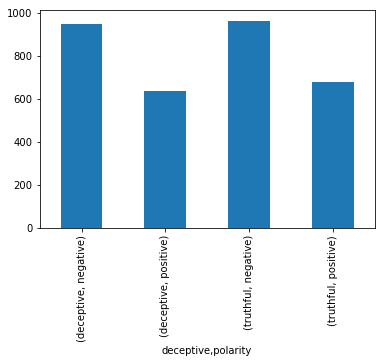
The first three rows in the dataset are shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **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. However, there is a statistically relevant difference between the positive and negative reviews. The bar charts below illustrate this examination.







### Natural Language Processing

Text Frequency-Inverse Document Frequency (TD-IDF)

First we apply a technique frequently used in data mining called TD-IDF. Each review is counted as one document of our corpus, and individual word frequencies are counted for each document. The frequency vectors for each document are combined into a matrix 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 is used to predict the binary class that the document belongs to: either deceptive or truthful. This matrix is used as the input for two machine learning algorithms: Linear Support Machine Classification and Logistic Regression.

Topic Modeling

The next portion of the analysis applies a natural language processing technique called topic modeling. Topic models group related words together to identify topics.

### The Models

With TD-IDF matrix input:

Logistic Regression

Linear Support Machine

Because the matrix we are using is a sparse matrix with the majority of the word counts for each individual word in the corpus being zero, most of the models require conversion of the matrix to an input with more “dense” data. For this reason, we will apply topic modeling to reduce the number of “topics” in the matrix from the total number of words in the corpus to 100, thus providing a more dense matrix to the models.

Topic Modeling

We will use the following six models from the scikit learn machine learning library for our predictive accuracy results comparison:

Logistic Regression

Gaussian Naïve Bayes

Support Vector Classification

Linear Support Vector Classification

Random Forest Classifier

Linear Discriminant Analysis

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. Gaussian 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. The non-linear version defaults to the rbm kernel which is optimized for polynomial rather than linear variables.

1. Linear Support Vector Machine (LSVM) is a method that uses points that best separate classes into two groups in order to predict classification.
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, and each prediction is summed up the winning prediction is chosen.
3. Linear Discriminant Analysis The purpose of linear discriminant analysis (LDA) is to estimate the probability that a sample belongs to a specific class given a data sample. Applying Bayes Theorem results in. The shared covariance matrix and mean vectors are estimated from the training data.

### The Results

TD-IDF Results

The prediction results for Linear Support Vector Classification are:

Accuracy score: 0.89375

Support Vector Classification

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | precision | | recall | f1-score | support |
| deceptive | | 0.86 | 0.94 | 0.89 | 77 |
| truthful | 0.93 | | 0.86 | 0.89 | 83 |
| avg / total | 0.90 | | 0.89 | 0.89 | 160 |

Logistic Regression

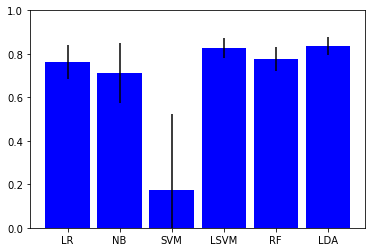
Accuracy score: 0.875

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| deceptive | 0.83 | 94 | 0.88 | 77 |
| truthful | 0.93 | 0.82 | 0.87 | 83 |
| avg/total | 0.88 | 0.88 | 0.87 | 160 |

Topic Modeling 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) |

**Topic Modeling Results Comparison**



### Results Analysis

## 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 binary classification use-case, the accuracy score is the most important evaluative tool.**

The most accurate classifications were made using our TD-IDF matrix as input to the Linear Support Vector Machine Classification algorithm. This model achieved an accuracy score of 89%.

The next most accurate classifications were made using our Topic Model vector as input to the Linear Discriminant Analysis Classification algorithm. This model achieved an 84% accuracy in recognizing deceptive reviews.

TF-IDF/Linear Support Vector Classification:

accuracy score: 0.89375

Runner-Up:

Topic Modeling/Linear Discriminant Analysis:

accuracy score: 0.845000

The booby prize goes to the Support Vector Machine algorithm on the topic model vector which predicted deceptive reviews less than 20% of the time. On examining why this model was such a dismal failure, it became apparent that this model uses a default kernel (rbf) which is designed to work not on a binary classification problem such as we have, but rather on a non-linear or polynomial use case. I re-ran the models using the linear (classification) version of the SVM model and it achieved an 83% accuracy. Clearly it’s important to choose the right version of the models for the type of problem we are trying to solve, whether it is a classification or regression analysis.

### 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>