**A Corpus of Truthful and Deceptive Hotel Reviews**

**By: Jennifer Adrianne I. Caballes and Christian Ted Ochoa**

**Dataset Description**

The dataset is a collection of truthful and deceptive reviews for 20 hotels in Chicago. It is taken from Kaggle, posted by Rachel Tatman [1]. The dataset contains five (5) different columns namely;

1. *Deceptive*: Classifies whether the review is truthful or deceptive.
2. *Hotel:* The column that contains the name of the hotel given the reviews.
3. *Polarity:* Classifies whether the sentiments are positive or negative.
4. *Source:* Contains the source of information.
5. *Text Review:* The actual review posted for the particular hotel.

Hence, A total of 1600 data is present and 80% (1280 data) will be used for training while the other 20% (3200 data) will later be used for testing.

**Frequency Matrix**

The table below shows how many data are present for a combination of feature. Thus, 400 truthful positive reviews from TripAdvisor , 400 deceptive positive reviews from Mechanical Turk , 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp , 400 deceptive negative reviews from Mechanical Turk

|  |  |  |
| --- | --- | --- |
|  | DECEPTIVE | TRUTHFUL |
| DECEPTIVE | 800 | 0 |
| TRUTHFUL | 0 | 800 |
| POSITIVE | 400 | 400 |
| NEGATIVE | 400 | 400 |

**Results**

**Insights**

**References**

Tatman, R. (2017, July 18). Deceptive Opinion Spam Corpus. Retrieved March 17, 2018, from <https://www.kaggle.com/rtatman/deceptive-opinion-spam-corpus/discussion>

Deceptive Opinion Spam Corpus Myle Ott. (n.d.). Retrieved March 17, 2018, from <http://myleott.com/op-spam.html>