NLP Techniques for Detection of Fake Reviews for Hotel Industry

Introduction

The impact of online reviews on businesses has grown significantly during last years, being crucial to determine business success in a wide array of sectors, ranging from restaurants, hotels to e-commerce. Some users use unethical means to improve their online reputation by writing fake reviews of their businesses or competitors. Detecting true and deceptive reviews has become a mandate for the hotel industry to survive the competition. This project proposes a feature framework for detecting fake reviews that has been evaluated in the hotel consumer domain.

About the Dataset

The dataset contains a collection of hotel reviews data. Each observation consists in one customer review for one hotel. A review is composed of a textual feedback of the customer's experience at the hotel and an overall rating. Variables in the dataset include:

Variable	Meaning
Hotel_Address	Address of hotel.
Review_Date	Date when reviewer posted the corresponding review
Average_Score	Average Score of the hotel, calculated based on the latest comment in the last year
Hotel_Name	Name of Hotel
Reviewer_Nationality	Nationality of Reviewer
Negative_Review	Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
Review*Total*Negative*Word*Counts	Total number of words in the negative review.
Positive_Review	Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
Review*Total*Positive*Word*Counts	Total number of words in the positive review.
Reviewer_Score	Score the reviewer has given to the hotel, based on his/her experience
Total*Number*of*Reviews*Reviewer*Has*Given	Number of Reviews the reviewers has given in the past.
Total*Number*of_Reviews	Total number of valid reviews the hotel has.
Tags	Tags reviewer gave the hotel
days*since*review	Duration between the review date and scrape date.
Additional*Number*of_Scoring	There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there.
lat	Latitude of the hotel
lng	longtitude of the hotel

Objectives of the Project

1. To understand the type of review shared by the customers.
2. Perform sentiment analysis on the reviews.

Methodology

Perform Sentiment analysis using the Natural Language Processing (NLP) techniques to extract emotions related to the raw texts. Analyze the customer reviews to determine whether they are positive or negative.

Use 'NLTK' library.

Challenge

For each textual review, predict if it corresponds to a good review (the customer is happy) or to a bad one (the customer is not satisfied). The reviews overall ratings can range from 2.5/10 to 10/10. In order to simplify the problem reviews have can be split those into two categories:

- bad reviews have overall ratings < 5
- good reviews have overall ratings >= 5

The challenge is to be able to predict the information using only the raw textual data from the review.