CSE 4062 - Semester 2021

Intro to Data Science and Analytics

Fake Review Detection

MARMARA ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ

Group 2 Members:

Aleyna BOZACI 150319630 - aleynabozacii@gmail.com

Bekir ÖZKAN 150319557 - bekirozkan9698@gmail.com

Merve YAYIN 150116051 - yayinm8@gmail.com

Metehan ERTAN 150117051 - metehan.ertan@hotmail.com

Mustafa USCA 150416054 - uscamustafa64@gmail.com

1- Project Description

Online product reviews are a fundamental part of the decision-making process for customers as well as vendors on e- commerce. Prior to purchasing services or goods, customers first review the online comments submitted by previous customers. However, these comments can be deceiving as they can be spam or fake. These misleading reviews can cause huge damages to company reputation and services or goods. This is a strong incentive for people to game the system and manipulate user sentiment by posting fake opinions or reviews to promote or to discredit some target products. Our project will tackle this problem of spam/fake reviews by developing a model, which could classify a given review as either fake or genuine, thereby helping to make more meaningful review information available to the customers.

We will be using CNN (Convolutional Neural Network), NB (Naïve Bayes), RF (Random Forest) algorithms. We will be improving these methods in order to achieve more efficiency and better results.

2- Dataset

2.1 - Deceptive Opinion Spam Corpus v1.4s. We will be using negative reviews for testing.

- 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

2.2 - We extracted a smaller dataset from YELP Dataset as it was too large. We are going to use 800 truthful and 800 deceptive reviews for the learning part and 200 reviews for the testing part.

For our project each column will be a word from review and each row will be that review. Column count changes with the length of review. Deceptive Opinion Spam Corpus v1.4s has 800 rows for learning and 800 rows for testing. For the Yelp dataset there are 1600 rows for learning and 200 rows for learning. Target attribute is truthful or deceptive as boolean.