Bekir Özkan 150319557
Merve Yayın 150116051
Metehan Ertan 150117051
Aleyna Bozacı 150319630

CSE4062
Introduction to Data Science and Analytics
Fake Review Detection

1. Topic

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.

2. Dataset

Deceptive Opinion Spam Corpus v1.4 consist of 4 parts.

- 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

The Yelp dataset is a subset of Yelp's businesses, reviews and user data. This dataset is opensource for personal, educational and academic purposes. Dataset is available as a JSON file. This data includes 608,598 reviews for restaurants. This dataset contains reviews from 5,044 restaurants by 260,277 reviewers. In this dataset, there exist 13.22% filtered reviews by 23.91% spammers.