**PRODUCT REVIEW CREDIBILITY ANALYSIS**

A Capstone Research Project Paper Abstract

**Faculty Research**

Zach Shim

**University of Washington**

August 2021

**Faculty Advisor:**

Dr. Min Chen

**Fake Review Detection on Amazon**

**Abstract**

Product reviews on web-based shopping sites have become a vital asset for online consumers. The reviews of a product can significantly sway an individual’s opinion about a product. Therefore, malicious sellers will deliberately manipulate reviews on their products to sway the opinions of potential buyers. Machine learning models can be leveraged against spammed or ingenuous reviews through pattern detection and analysis.

The faculty research I worked on involved scaling a legacy project that detects fake reviews from Amazon. A collection of data analysis algorithms worked in tandem to form a classification system that takes the text of a review and basic information about its reviewer as input and outputs a ‘fakeness’ score that indicates a product’s credibility. The system uses three methods for analysis: detection of duplicate reviews, detection of incentivized reviews, and detection of anomalous reviews.

I expanded on this existing framework to make the system web based, more efficient, and easier to scale. The product was a website with a refined web interface that allows users to dynamically analyze new data by pasting links to an Amazon product’s webpage. The system then scrapes relevant data from the product page and feeds this information into pre-trained machine learning models for analysis. The pipeline of the link pasting page follows a similar structure to the original methods of analysis: similarity detection, sentiment detection, and anomaly detection.

I experimented with a number of different libraries and machine learning models. I used an unsupervised model for similarity detection, which featured a mix of natural language processing, TFIDF vectorization, and latent semantic indexing. A cosine similarity heuristic was used to determine the similarity between all pairwise vectors in the corpus.

This application can be a useful asset to consumers that are unsure if a product’s reviews are fake. Using this application can give them confidence that their purchase is free of malicious reviews.