Adita Jha, Jim McCarthy

University of San Diego

AAI-510-03 Machine learning: Fundamentals and Applications

Professor David Friesen, Ph.D

19 June 2025

# Abstract

# Introduction

Needs to be re-written.

The purpose of this project is to develop a set of ML models that will predict the expected rating for a review given a set of reviews and their associated ratings. Two publicly available databases will be used. The first is a sample of Amazon’s ratings database, and the second is from Yelp’s ratings database. Both databases have a large training set, with the Yelp training set the larger of the two. The project was selected, because the output for the model that we create has a limited output domain, a discrete set of integer values between 1 and 5.

A combination of a number of different techniques were used in this project. A supervised learning linear regression model and the use of Natural Language Processing (NLP) for building context on the review text is likely needed. We will also investigate the use of vector embedding to help build relationships between words with similar meanings or usage patterns. It is likely that we will start with Gensim or spaCy for vector embeddings, as they provide some pre-trained models. Since we are not using domain-specific vocabulary, it is not necessary for a new custom embedding to be created.

It is likely that the data will need to be pre-processed to remove punctuation and stop words, convert to lowercase, convert text to numbers, and correct spelling. There are also other processing steps that need to be performed to reduce the dimensionality of the dataset, though what these are we will need to determine.

# Data Used

Amazon Review data

* Fields used

Yelp Review data

# Data Analysis

Graphs used

Information gleaned from graphical representations

Tables used for analysis

## Key Insights from Data Visualization

Insights gained from data analysis

Imbalance of data and ratings

Review text was short

# Feature Selection

Field selection

* Review text
* Summary
* Is review useful

Target feature

* Ratings, range and expected values.

Obfuscated and time related fields ignored

# Data Preparation

Data preprocessing

* Lower casing, punctuation removal, stop word removal
* Null removal
* Removal of rows with missing rating, summary, or review text
* Casting of string values to integers as needed

Lemmatization

Down sampling to address imbalance of ratings

Split into train / validation / test sets.

# Modeling

Text vectoring representation

Use of TF-IDF + Ridge regression

Use of TfidfVectorizer to convert text to feature vectors.

BERT + Random forest

# Performance Evaluation

How well did model predict rating from review text.

Confusion matrix summary

Low and high value prediction compared to mean

Regression to mean behavior, and how to avoid

# Deployment

Main parts of deployment

U/I inputs and outputs

Database storage

Process flow

Internal versus external customer

# Discussion

Imbalanced data, and how to counter bias

Model performance comparison

Tuning parameters for different models

# Conclusions

What was found and how the models performed

Which performed best

Future work that could be explored

# References

# Appendix A