Rodrigo Rodriguez

07/03/2024

Applied Data Science

DSC680 (SUMMER)

# Week 8 Milestone1: Automating Online Cookies Classification to Mitigate Legal Risk

## Topic

My last project involves building a classification model to automate the classification of online website cookies. Cookies must be classified into one of the following categories:

- Essential Cookies
- Performance Cookies
- Functional Cookies
- Targeting Cookies

## Introduction

This is an active project at my employer, waste Management, and has been shared with permission. WM is a Fortune 200 company that provides waste and recycling management services. WM has thousands of online cookies across all their web domains. With recent online privacy laws like, California's CCPA, a legal risk exists by leaving all cookies active. Online cookies must be classified in order to give website users the choice of opting in or out of certain cookies.

## Problem Statement

The goal of this project involves taking natural text data related to online cookie descriptions and using a supervised learning dataset to train a classification model. The model will predict the classification of online cookies. The intent is to test and train a classification model to scale the efforts of classifying online cookies across all of WM's web domains.

## Datasets

The dataset is obtained internally from my employer (WM) with permission. The dataset is a structured dataset with natural language text as the primary predictor. This data will be used to build a NLP classification model to predict the appropriate category for each cookie based on the cookie's description.

## Methods

The methods I will use follow the established Machine Learning phases. These include the following:

- **EDA**: Visualizing and summarizing the data to identify patterns and relationships.
- **Data Preparation**: Cleaning the dataset, handling missing values, and encoding categorical variables.
- **Data Processing**: Transforming the data if needed, vectorization, feature engineering, etc.
- **Modeling**: Applying ML modeling techniques.
- **Model Evaluation**: Measure the effectiveness of the models, MAE, MSE, etc.
- **Interpreting Results**: Analyzing model coefficients to understand the impact of different variables.

For this project I plan to test a Multinomial Naïve Bayes classification model. Naïve Bayes is an appropriate model for this project because of the model's tendency to learn patterns in the training data well. Since contextual meaning in cookie descriptions tend to not change, I believe that a Naïve Bayes model will be highly accurate (Yiu 2019). Additionally, a Naïve Bayes Classification model can handle multiple variables.

## Ethical Considerations

All cookies must have accurate descriptions for the model to be able to learn and classify cookies into the correct categories. The risk of inaccurate cookie descriptions could pose an ethical concern as cookies with a false or missing description could inaccurately be classified as essential and go against the company's privacy policy.

## Challenges / Issues

The primary challenge is the limited dataset available for training purposes. Initial EDA revealed that there is only 188 records available in the dataset. Additionally, class imbalances were observed in the dataset. The data will have to be properly split to ensure proper class representation in both training and test sets.

## References

I plan to reference both academic and online sources that explain the benefits of the models that I plan to test. Additionally, any sources that explain online cookies will also be useful to reference.

## Sources

Yiu, T. (2019, October 24). *Understanding The Naive Bayes Classifier*. Towards Data Science. Retrieved May 24, 2024, from https://towardsdatascience.com/understanding-the-naive-bayes-classifier