CIS 579 – ARTIFICIAL INTELLIGENCE

SENTIMENTAL ANALYSIS ON CAR REVIEWS

WINTER SEMESTER – 2024

Team Members –

1. Navaneeth Gopireddy
2. Rohith Karre
3. Deepthi Subramanyam
4. Anish Kolaparthi

Abstract:

We are now witnessing a paradigm shift in how technology is integrated into every field, shaping our daily lives and interactions. One of the most remarkable transformations in the reviews has revolutionized the way a company can make improvements in the respective modules. The abundance of online car reviews provides a treasure of valuable insights into consumer sentiments and preferences within the automotive industry. Car manufacturers, dealerships, and other stakeholders can leverage sentiment analysis techniques to extract meaningful information from these reviews, ranging from identifying areas for product improvement to fine-tuning marketing strategies and monitoring brand reputation. By systematically analyzing and interpreting the sentiments expressed in car reviews, businesses can gain a competitive edge by better understanding customer needs, enhancing product offerings, and ultimately fostering greater customer satisfaction and loyalty.

Problem Statement:

The objective is to analyze the sentiments expressed in these reviews to identify areas for product improvement and derive actionable insights that can direct the improvement of the automobile product by comprehending the feelings expressed by customers. to classify the reviews as either positive or negative.

Background and AI Type:

Supervised learning is a type of machine learning where the model is trained on a labeled dataset, meaning each input data point is paired with the correct output. During training, the model learns the mapping between inputs and outputs so that it can make predictions on new, unseen data. This approach is commonly used for tasks such as classification and regression.

Neural networks are a class of models inspired by the structure and function of the human brain. They consist of interconnected nodes, called neurons, organized into layers. Each neuron receives input, performs a computation, and passes its output to the next layer. Through a process called backpropagation, neural networks learn from data by adjusting the weights of connections between neurons to minimize the difference between predicted and actual outputs.

We used both Supervised models and Neural Networks. The latter also comes under Supervised in a way as we used training data with labels to build the network.

Dataset Description:

The Kaggle repository provided the dataset that was used in this project. We used the "Edmunds-Consumer Car Ratings and Reviews" dataset specifically. This dataset consists of CSV files with reviews for fifty different brands and the models that go along with them. To simplify our study, we combined all the CSV files into one cohesive dataset. After the process of consolidation was finished, the dataset included 284,715 records that were made up of different brands and models. The variables of our dataset are:

- Review date: date of the review posted.
- Author Name: Name of the person who posted the review.
- Vehicle Title: Model and brand of the car.
- Review Title: Title of the review.
- Rating: Rating of the review.
- Review: Actual review for the car.

Provided below the snippet of our dataset.

	Unnamed: 0	Review_Date	Author_Name	Vehicle_Title	Review_Title	Review	Rating
0	0	on 01/24/17 17:50 PM (PST)	jose Linares	1998 Mazda B-Series Pickup Regular Cab B2500 S...	My first pickup.	Excellent unit overall.	5.000
1	1	on 09/14/09 14:41 PM (PDT)	Patersg	1998 Mazda B-Series Pickup Regular Cab B3000 S...	Old Faithful	Had the 3.0L V6 regular cab for several years...	3.750
2	2	on 04/10/09 04:03 AM (PDT)	The Pro	1998 Mazda B-Series Pickup Regular Cab B2500 S...	dreamy, dependable truck	I have had this truk a very long time, I own...	4.625
3	3	on 09/03/05 10:22 AM (PDT)	swcejk2002	1998 Mazda B-Series Pickup Regular Cab B2500 S...	Amazing	Amazing small truck. I put A LOT of miles on ...	4.750
4	4	on 08/31/04 00:00 AM (PDT)	dr.dimento	1998 Mazda B-Series Pickup Regular Cab B3000 S...	bad brakes	I cant keep brakes in this truck.Three\nsets ...	3.250

Link to dataset: https://www.kaggle.com/datasets/ankkur13/edmundsconsumer-car-ratings-and-reviews

Data Acquisition Challenges:

Initially, we faced challenges with collecting the data since our dataset is scattered among multiple .csv files which we unified into a data-frame using efficient processing techniques.

Data Processing:

Before proceeding with our model processing, we had to do some data pre-processing. We divided our methodology into two parts – Sentiment polarity and additional features.

Sentiment Polarity is assigning polarized values for each word. We have used NLP techniques for assigning values. The 4 techniques are casing, removing stop words, stemming, and polarizing.

1. Casing**:** Refers to process of converting text into consistent case, such as converting all the letters into lower case or upper case, removing white spaces and special characters too. For our project, we have converted the review into all lowercase letters. This helps in standardizing the text for further analysis.
2. Removing Stop Words**:** Stop words are common terms in a language, such as "the," "is," "and" etc., that frequently have little to no significance in a particular context. Eliminating stop words from the text can assist tasks become more accurate and efficient by concentrating on the more meaningful terms.
3. Stemming: Removing the tail words to find out the root word. This helps in decreasing the dimensionality of the lexicon and treating many variants of the identical word as the identical token, hence enhancing text analysis tasks such as search or classification.
4. Polarizing: converting the reviews into polarized values using textblob so that they can be used in modelling. It involves determining the emotional polarity or sentiment expressed in the text, which can be valuable for understanding opinions, attitudes, or reactions in textual data.

The additional features that obtained are count words, unique count words, count letters and mean word length of the review.

For our target sentiment, we have chosen that if the sentiment polarity is greater than 2.5, then it is classified as positive sentiment or if polarity is less than or equal to 2.5 then it is considered as negative sentiment.

Model Overview:

Logistic Regression:
It is a supervised learning algorithm used for binary classification tasks. Logistic regression was applied to predict the sentiment of car reviews. It models the probability of a positive or negative sentiment of a review using a logistic function.

Decision Trees:
These are versatile algorithms for both classification and regression tasks. Decision trees are known to interpretable and can handle both numerical and categorical data.
They construct a flowchart-like structure where each internal node represents a decision based on features, leading to leaf nodes that correspond to class labels in this case if a review is positive or negative based on the threshold.

K-Nearest Neighbors (KNN):
It is a non-parametric and lazy learning algorithm used for classification and regression tasks. It assigns labels to data points based on the majority class among their k nearest neighbors. In this project, KNN can be used to classify car reviews based on similarity to other reviews in the dataset.

Support Vector Machine (SVM):
It is a powerful supervised learning algorithm used for classification and regression tasks. It finds a hyperplane that best separates different classes of data points with the maximum margin. SVM can handle non-linear decision boundaries using kernel tricks, making it suitable for complex car review classification task.

Random Forest:
It is an ensemble learning method that builds multiple decision trees and combines their predictions. It improves upon decision trees by reducing overfitting and increasing accuracy. In this project, random forest can be effective for analyzing car reviews by capturing complex interactions between different review features.

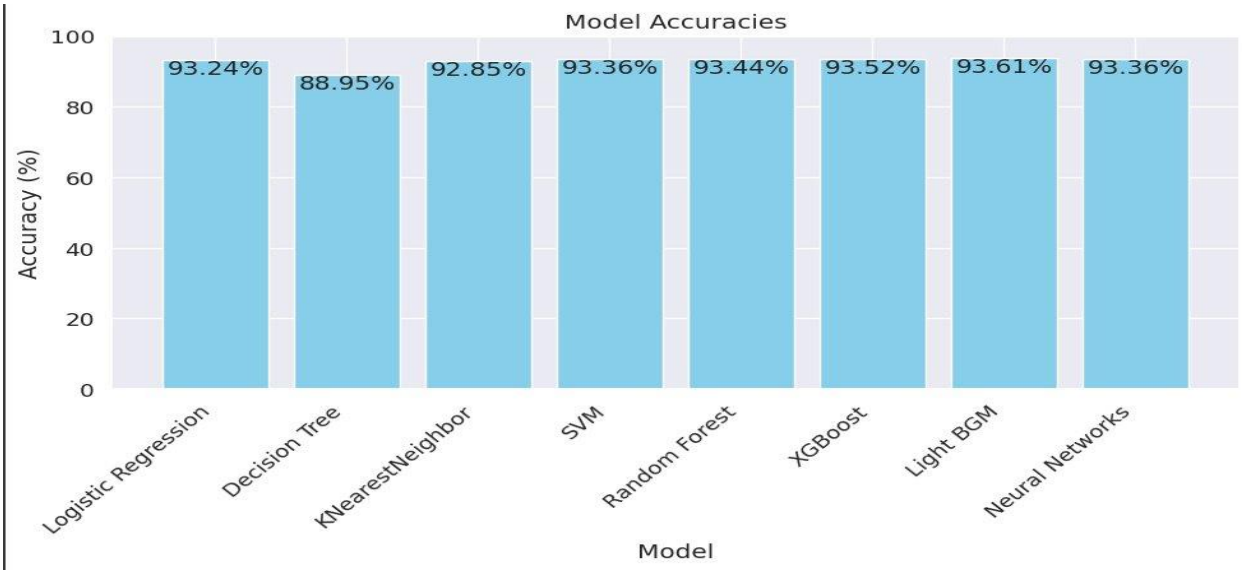
XGBoost (Extreme Gradient Boosting):
It is an efficient and scalable implementation of gradient boosting. It builds an ensemble of weak learners, multiple decision trees in a sequential manner to improve predictive performance. XGBoost is known for its speed and accuracy, making it suitable for large-scale tasks like car review analysis, where the amount of reviews can be many. The algorithm sequentially builds models to minimize overall prediction errors. XGBoost uses pre-sorted algorithm & Histogram-based algorithm for computing the best split. Each model focuses on correcting the mistakes made by its predecessor by giving higher priority to misclassified instances, gradually improving the overall accuracy of the ensemble. Feature importance can be a valuable metric to understand which features are contributing the most to the model's predictions. XGBoost provides built-in methods to calculate and visualize feature importance.

Light GBM (Light Gradient Boosting Machine):
It is another gradient boosting framework that uses a novel tree learning algorithm to achieve high performance and efficiency. It optimizes memory usage and training speed, making it suitable for handling large datasets. Light GBM can effectively capture intricate patterns in the data. LightGBM uses Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) where it bundles mutually exclusive features by focusing on larger gradient instances and perform automatic feature selection, enhancing the boosting process finding a split value.

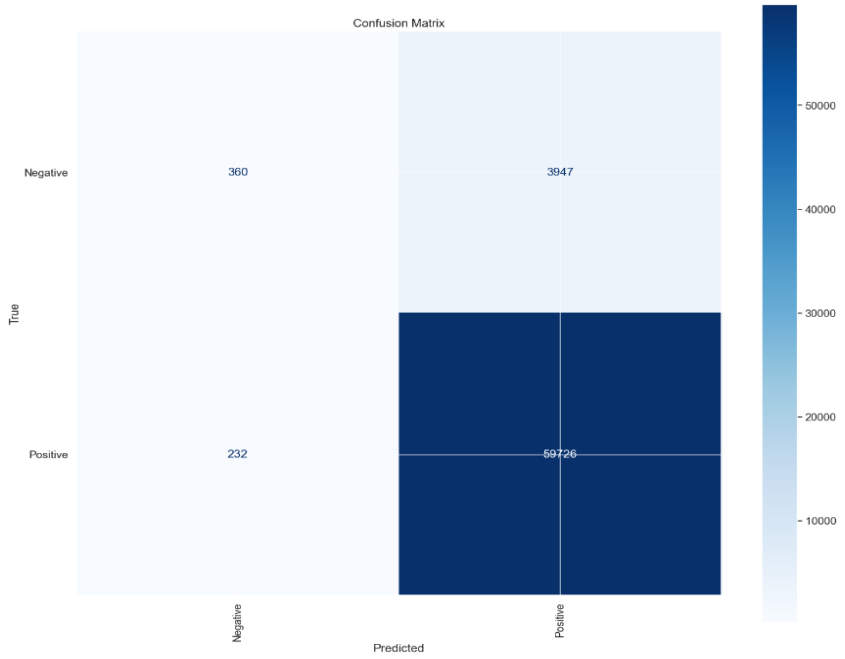
Neural Networks:
Neural networks excel at learning complex patterns from unstructured data and can automatically extract features for sentiment analysis or review classification tasks.
Here we used a Basic ANN with the input layer and a hidden layer with ReLU activation and an output layer with Sigmoid activation.

Results:

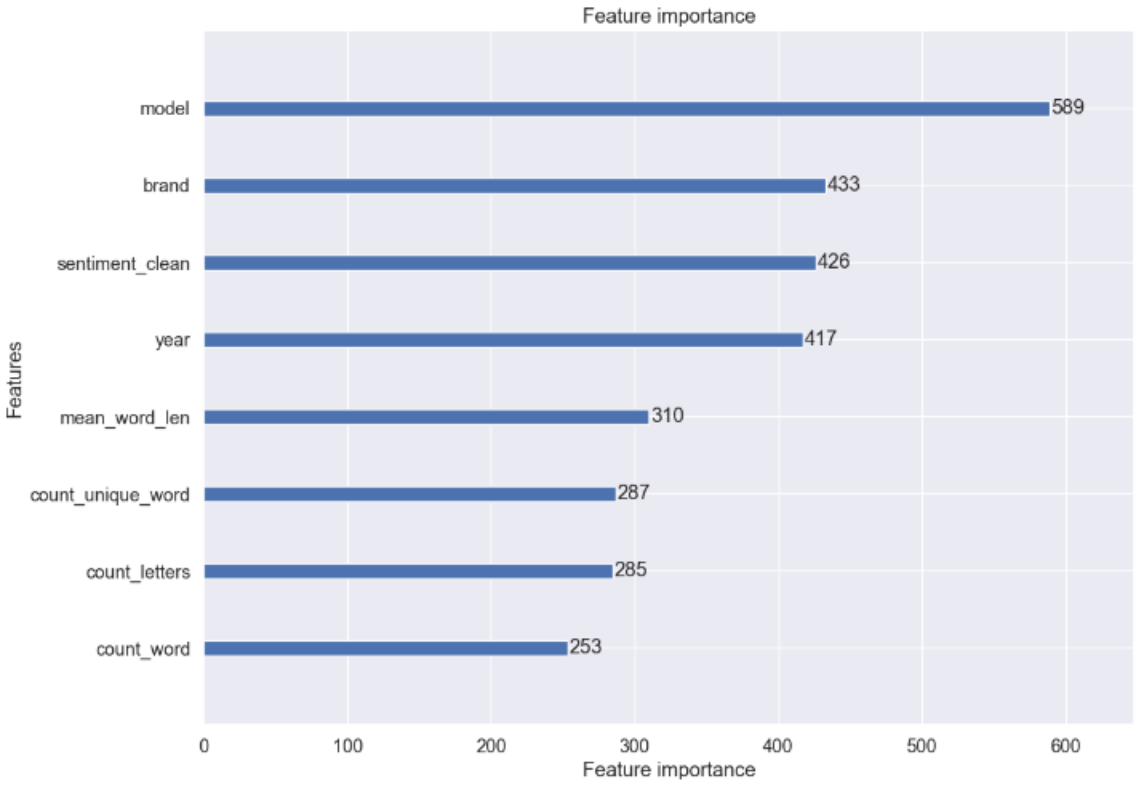
In our project on sentiment analysis of car reviews, we conducted testing on 30% of the dataset, totaling 65,435 records. Among these, there were 4,909 negative sentiment records and 59,353 positive sentiment records. Accuracy served as the primary metric for model evaluation. Notably, we observed varying levels of accuracy across different models. Logistic regression yielded an accuracy of 93.28%, Decision Tree at 88.95%, KNN at 92.85%, SVM at 93.36%, Random Forest at 93.44%, XG Boost at 93.52%, Light GBM at 93.61%, and Neural Networks at 93.24%. Among these models, Light GBM showcased the highest accuracy at 93.61%, making it the best-performing model for sentiment analysis in our project.



The finalized model based upon the accuracies is Light GBM with an accuracy of 93.61%. Below is the confusion matrix of the model which illustrates its performance.



Feature ranking refers to the process of determining the importance or contribution of each feature towards the model's predictive performance. This is crucial for understanding the impact of different features on the model's decisions and for feature selection or dimensionality reduction.



The sentiment_clean variable stands at third place in the importance indicating that it is one of the critical variable in predicting the sentiment of the review.

Conclusion:

Project Impact:
The results from our best-performing model show us which car models of a particular brand offer the customers the highest satisfaction and which do not. These kinds of effective analysis can be very beneficial when it comes to designing products for a particular set of customers. Our sentiment analysis model can also be used in various fields where we want to know customer satisfaction and let customers guide the organizations in shaping better products for the future.

Future Plans:
Future enhancements of our project include procuring a more nuanced understanding of the customer reviews by adding more classes to our target variable such as semi-positive sentiment, neutral sentiment, and semi-negative sentiment. These kinds of additional details can help us get a more detailed analysis of reviews.

Summary:
The most interesting part of developing our model was using various natural language processing techniques to process the text and get the emotional depth and intensity exhibited by various customers. We are very proud of the fact that we succeeded to a great degree in yielding the proper sentiment expressed in each review.
Project Overview:
Our project on sentiment analysis of car reviews has provided valuable insights into the effectiveness of various machine learning models in accurately classifying sentiments. All the models performed similarly well, but Light GBM was the most efficient of them all because of its efficient feature bundling and gradient one-sided sampling. This shows that it is reliable and appropriate for the job at hand. Overall, our results highlight how crucial it is to choose the right model for sentiment analysis tasks depending on the needs and features of the dataset.

References:

1. https://towardsdatascience.com/twitter-sentiment-analysis-on-car-brands-c13d449eb9fc
2. https://www.ijraset.com/research-paper/automobile-reviews-using-machine-learning-techniques
3. https://arxiv.org/pdf/2101.06053.pdf