NATURAL LANGUAGE PROCESSING

CSCE 5290 – SECTION -003

PROJECT PROPOSAL

TEAM: PROJECT GROUP 20

PROJECT TITLE:

PREDICT CAR PRICE USING NATURAL LANGUAGE PROCESSING

Team Members:

1. Name : Vishnu Vardhan Reddy Sudireddy
   ID : 11642773
   Email : vishnusudireddy@my.unt.edu

1. GOALS AND OBJECTIVES:

1.1 Motivation :

The use of natural language processing to predict the price of the car has not been implemented so far. In general, price of a car with different models can be viewed by the customers on website of the manufacturer or car seller. A customer who is willing to sell his car has to go the website and fill the forms with the details of the car such as make year, manufacturer and model of the car along with the milage needs to be provided for the further assessment of car price. This approach of finding the car price needs software development team to create forms and maintain which is a costly practice.

1.2 Significance:

Integrating natural language processing with the voice assistant can cut down the development and maintenance cost and this will reduce the time of the customer in finding the car price as he can directly ask the voice model to predict the price of the car. As the time of finding price of the car is much faster and easier compared to filling the forms, customer will get the response as soon as he asks the question. So, he/she can take the decision at the instant which will subsequently help in raise in sales of the seller.

1.3 Objective:

The main objective of this project is to build a car price prediction system that can help the customers and car sellers. With the implementation of this system customers can directly query the system with the car's make, model, manufacturer and milage in natural language and can obtain the response immediately. This will reduce the price of building the

1.4 Features:

1.4.1 Data Gathering:

Data about the car sales, prices, year, manufacturer, milage needs to be gathered from different sources like carsguru, carvana or Kaggle.

1.4.2 Data Preprocessing:

The data gathered may have irrelevant data at some instants. The irrelevances can be null values, negative values in the prices of the car and duplication of data. The preprocessing can also have the steps – removing stopwords, lemmatization etc. After cleaning the data, dataset get rid of any missing values and combines data from different sources into a single dataset.

1.4.3 Analysis of Data:

Analysis of data using various data visualization techniques will be performed to understand the underlying patterns in the data. We may use matplotlib, seaborn

for visualization of data. We will also use pandas and Numpy for the analysis of data.

1.4.4 Model Selection:

We may use different machine learning models like Radomforest, Decision trees and various other models to predict the price of the car. We will also use libraries like NLTK, spaCy, Gensim and various other libraires to perform text processing, text tockenization, stopword removal and lemmatization. Based on the performance metrics of various models we try, we will select the model with best accuracy.

1.4.5 Speech to text conversion

Python speechrecognition library will be used in this project to convert speech to text. The converted text is then processed and analyzed to identify the car model, make, manufacturer and milage.

1.5 Deliverables:

We will be building a fully functional model such that it can be reused and easily modifiable based on requirement.

1.6 Uniqueness:

In general, the value of a car is analyzed by various parameters of the car in a form in a website or using a manual inquiry. Finding the value of the car by providing speech as input is not evident in the market yet. So, I can say that my project is unique. This system can be reused in different retail areas like bike retailers, resale of electronic refurbished or renewed products.

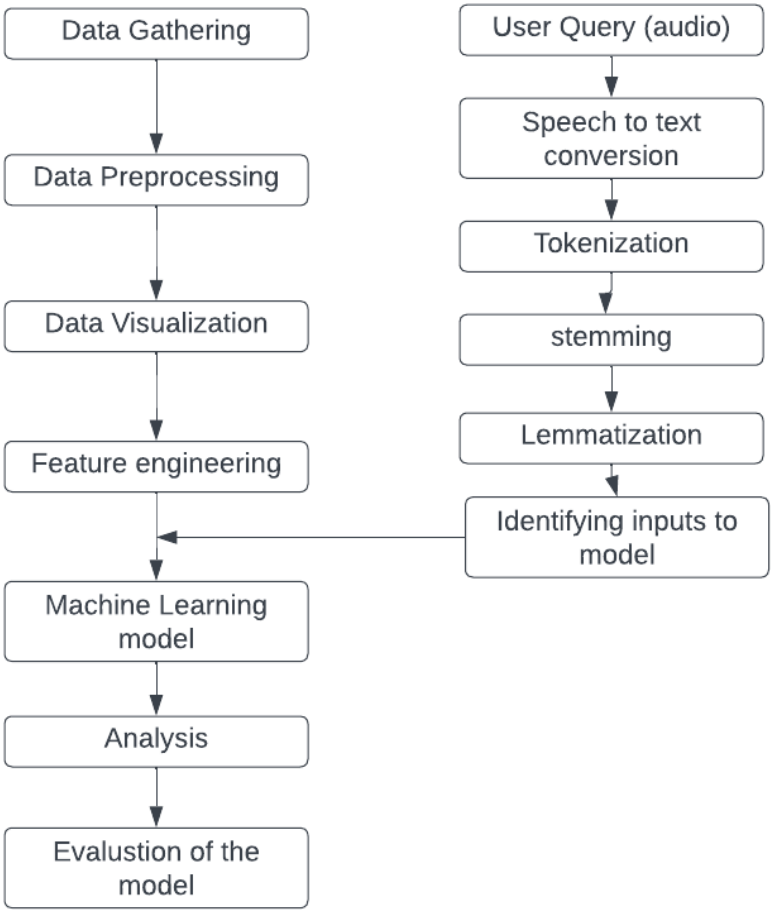
1.7 Milestone:

Milestone	Week	Task
1	6	Data gathering and Data Preprocessing
2	8	Model Building
3	10	Speech to text conversion and text processing using NLP
4	11	Evaluating the models
5	12	Model Tuning and Visualizations
6	13	Improvements if possible

1.8 Technical Features of the Project:

I will be using python and its libraries in build this project. The libraries that I may use in this project are Numpy, Pandas, and NLTK, GENISM, spaCy, matplotlib, seaborn, performance metrics, Tokenization, Stemming / Lemmatization, Exploratory Data Analysis, data Visualization methods.

2. VISUALIZATION:



Git Link: https://github.com/vr-23/natural_language_processing

REFERENCES:

[1] https://www.kaggle.com/

[2] https://datasaur.ai/blog-posts/nlp-speech-recognition

[3] https://ieeexplore.ieee.org/document/9912940

[4] https://towardsdatascience.com/used-car-price-prediction-using-machine-learning-e3be02d977b2

[5] https://www.ibm.com/topics/natural-language-processing

[6] https://www.educative.io/answers/preprocessing-steps-in-natural-language-processing-nlp

[7] https://datasaur.ai/blog-posts/nlp-speech-recognition

[8] https://journals.sagepub.com/doi/abs/10.1177/0265532210364405?casa_token=Gi5jcRnEPcwAAAAA:BwujUSvD0pxJdPwsZa4ASRWPIaaFUbvwZjOBAKpC1GFHjqVAXmq0CaLQjgbI9DatFrQvJfPl7UdZgQ

[9] https://ieeexplore.ieee.org/abstract/document/9987967

[10] https://sol.sbc.org.br/index.php/eniac/article/view/9300