**REVIEW DATA ANALYSIS ON ELECTRIC VEHICLES USING NLP MODELS**

**TABLE OF CONTENTS**

[Introduction 3](#_Toc147330278)

[Discussion 3](#_Toc147330279)

[Conclusion 19](#_Toc147330280)

[Reference list 20](#_Toc147330281)

# 

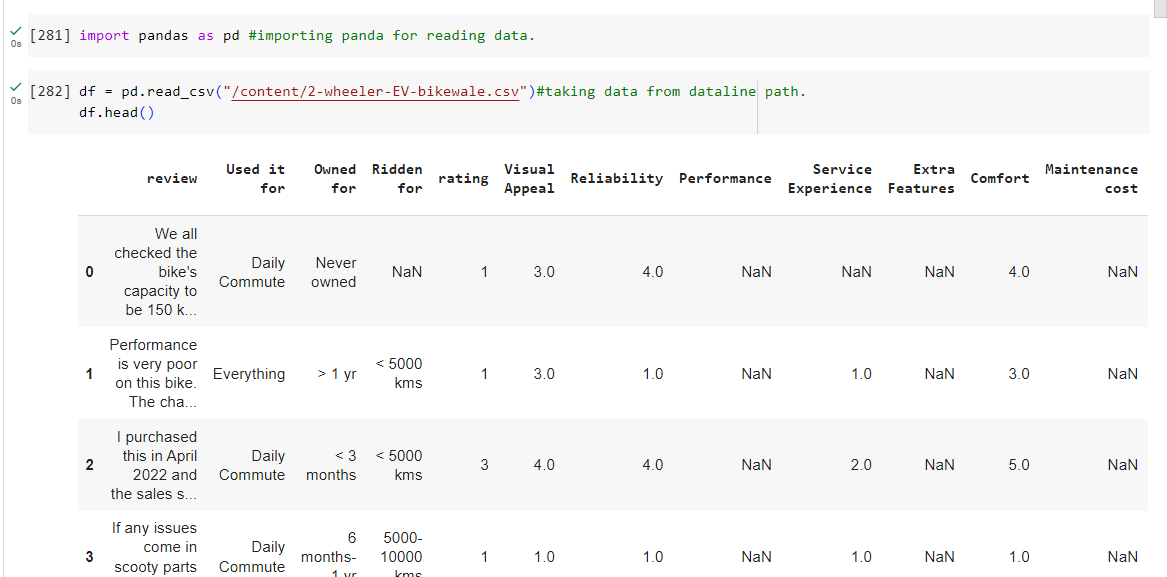
# Introduction

This assignment has a significant focus on the review of electric vehicles. By analysing the study trying to predict the company's insights and by the review analysis company can expect their next step in business. With the help of NLP techniques and data model training the company can predict the sentiment of all reviews. By those collective review parameters probability, they can understand the business trending and new prototype's nature and future scope.EV is a very new product in a competitive market world, to calculate the market for them this approach is very wise and efficient so that companies can understand the public emotions toward the products and the upgradations for the products which are needed for the prototype which is already in progress. By using Frequency Analysis, and Sentiment Analysis on the given data set of the EV bike this report will determine the basic emotions and mindset towards the EV in general.

# Discussion

This project is based on the sentiment analysis of the company product’s review. Here the company's product is EV bikes. EV is a very new product in a competitive market world. With this insight and the probability of the reviews, the company can understand the nature of their new products and their scope in the market. Here the model of analysis is based on 1. Frequency Analysis 2. Sentiment Analysis.

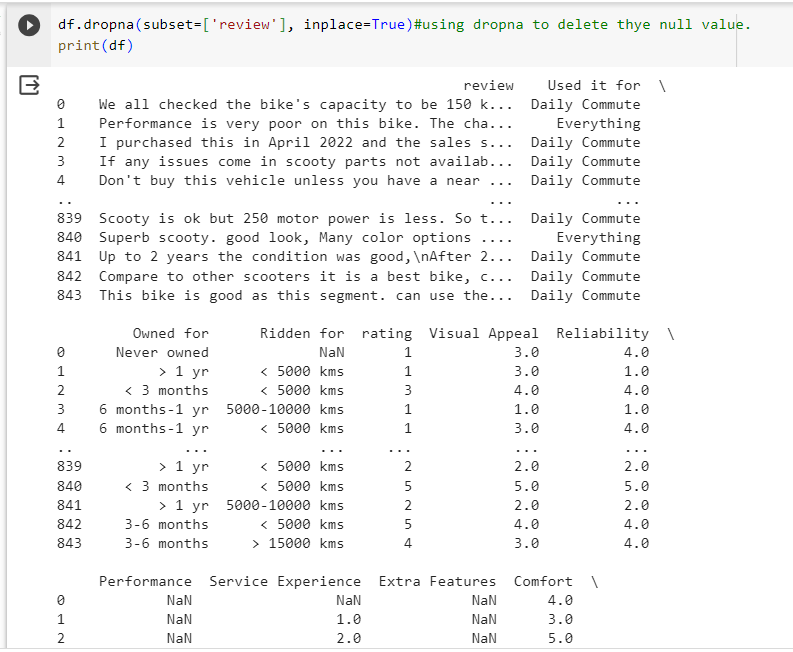
NLP stands for Natural Language Processing, AI has many subfields. This particular area of AI deals with the interaction between humans and computers. A computer is a complex mathematical summary optimization system that only understands 0 or 1 as an input sequence. Human language is very difficult to interpret and derive meaning from. NLP provides access to various functions to decompose human input text and help the computer understand its meaning. It involves developing models and using already trained models to help computers work with text and language data through various language-related tasks. NLP invokes text classifications through which a machine can label the data as sentiment analysis or other different analyses. Name Entity Recognition is a process through which a machine can understand objects and entities such as names and places. Machine translation helps the machine turn a text from one language to another. By speech recognition machine can understand speech and perform the command according to the giver’s orientations. Text generation is a very useful process, machines can generate only random text through continuous searching and the previous NLP task optimizations. Sentiment analysis has been used in this analytical data analysis of the report. Through this process, machines can detect the emotional tone of the text data this process will help to understand the reviews like customer feedback, market analyses, and social media analysis(Alonso *et al.* 2021). NLP gathered all techniques from computer science, deep learning, machine learning, and linguistics. This process involves some preordered steps like parts of speech tagging, tokenizations, semantic and syntactic analysis, and statistical rules-based trained modeling to make sense of the text. Customer review markets feedback to research and social media monitoring emotion analysis leads the business to understand how the market works and how the customer reviews their services and products (Zunic *et al*. 2020). This helps the company to improve the prototype on which those reviews are taken place. Sentiment mentioned on social media can help to understand the negativity and positive data to analyze particular situations and helps to manage the reputations and crisis response. In this process, the first requirement is to understand the basics and gather a data set that is relevant to the subject (Li, W *et al.* 2022). Here the data set is based on EV bikes, two two-wheelers. Fetching the data set in the required platform in this case that is Google Colab platform.

****

**Figure 1: Importing data set in Colab**

(Source: Self-created in Google Colab)

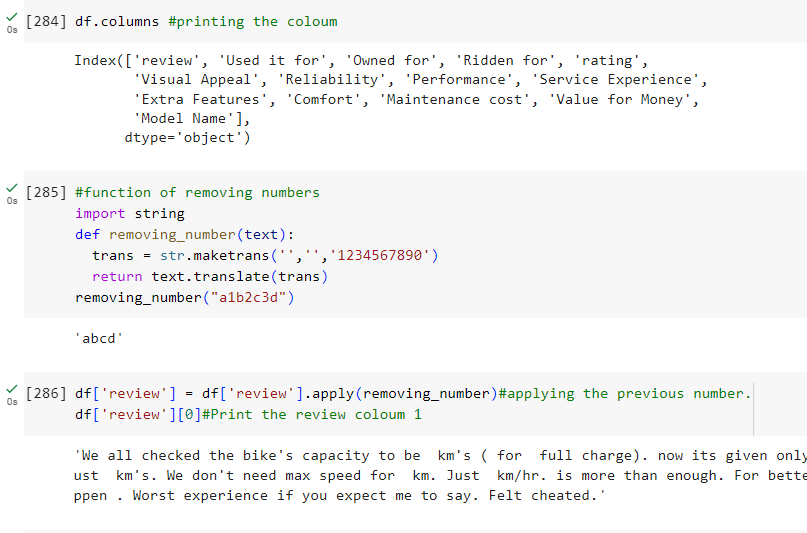
First import Panda’s package to read the CSV data from the file in the platform. Then the second step will occur, the data cleaning. Here first data cleaning took place by multi-processed method.

****

**Figure 2: Cleaning null value from the data**

(Source: Self-created in Google Colab)

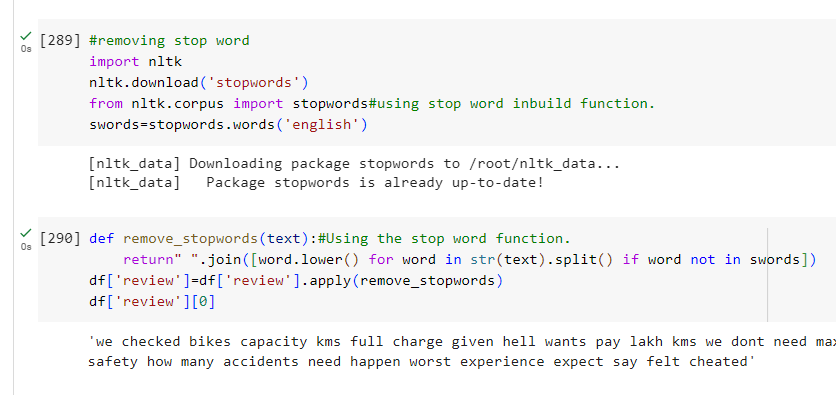
In this data-cleaning process, the first step is to clean all null data from the dataset. That will help in the consecutive steps in the code. Data cleaning is very important in this analysis assignment(Chakraborty *et al*. 2020). That cleaning process is based on the particular data which is given in the form of a CSV file here.

****

**Figure 3: Removing numbers from the data to clean**

(Source: Self-created in Google Colab)

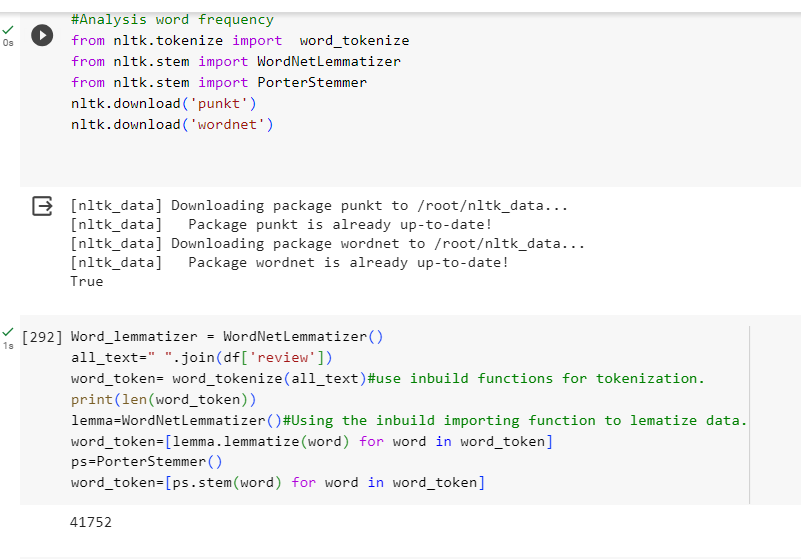
After cleaning the data display all columns for future reference. Then create a function of removing the numbers from the strings of reviews for data cleaning(Abualigah *et al*. 2020). Here are the many-step methods followed for data cleaning by which data will get ready for future endeavors.

****

**Figure 4: Removing stop word from the data**

(Source: Self-created in Google Colab)

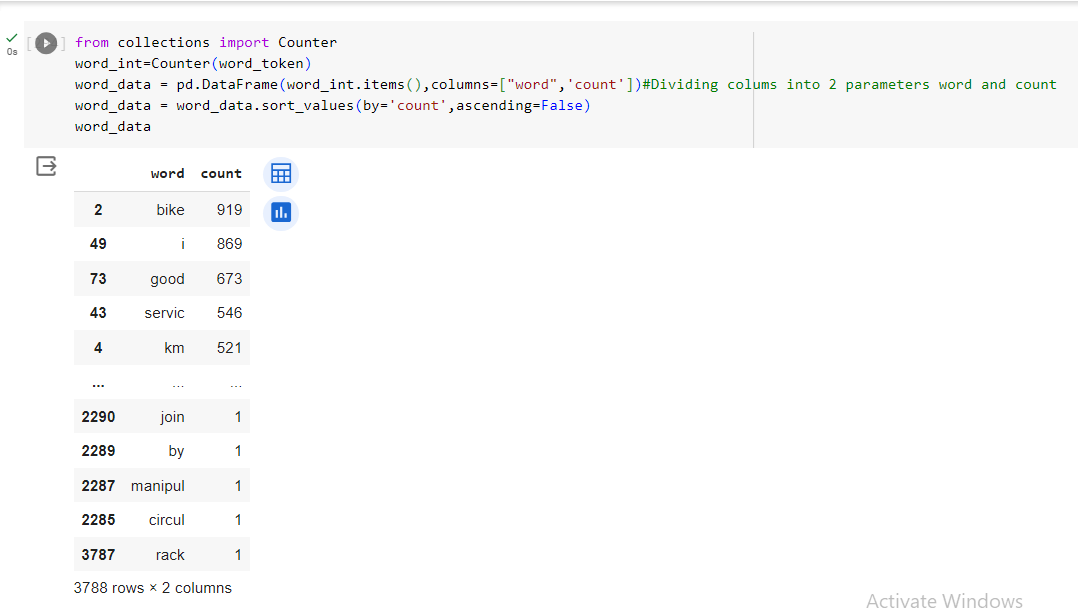
Stop words mean common words in the language set like the, is, in those are the often escape from the data based on the text while Natural language processing and the tastes like text management and analysis. The real cause of eliminating stop words is to diminish the dimensions of the data and improve the effectiveness of the analysis and text processing. By tokenizing the process of dividing the text into words and tokens NLP libraries or expressions will assign the tokes in the text. By NLTk libraries the data will remove any words that show the list of stop words, on which the company can do further comprehensive work.

****

**Figure 5: Data frequency analysis and tokenizing the data**

(Source: Self-created in Google Colab)

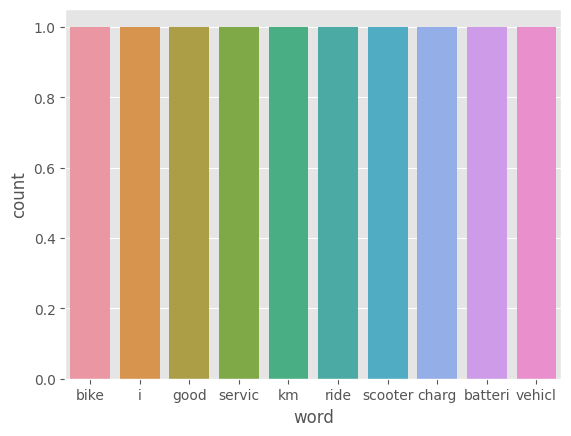
Word Frequency means how many times a word occurs in a given text or in a collection of data based on text. In the code, the main agenda was to calculate the most repetitive word from the reviews and based on that create the next steps. Topic modeling and text classifications by this process of word frequency features are usually used as the input of models based on machine learning and text classifications. By word lemmatization machine can reduce the words that are based on the lemma. This lemmatization process is a very useful process in that machib\ne standardizes the words from the different tense forms as the same word.

****

**Figure 6: Division of the review columns**

(Source: Self-created in Google Colab)

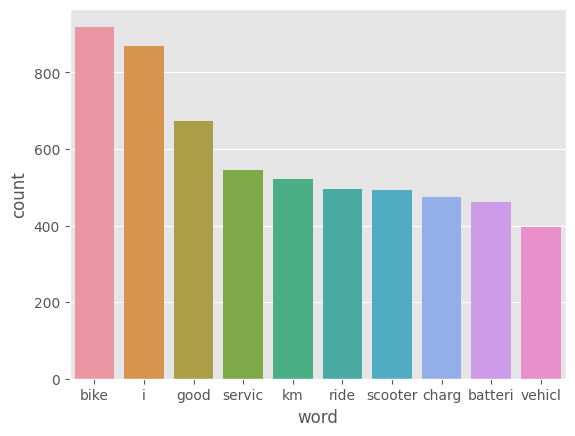
In this process with the help of the collection library from the data set count the words which occur how many times. Here as bike, the result is very prominent that bike good service keywords are used in significant number of times. This tells us that the data is ready to pursue from the model we want to give. This process is very useful for visual purposes and determining the repetitive keywords, This will help the code to determine the repetitive keywords that are very frequent in the data set and pot the probability against the respective manner and by that predict the future steps. This process is called MONO gram where we see the single word string from the whole dataset and recognize the important keyword from the data set which helps to understand the future scope of the dataset.

****

**Figure 7: Count plot of the data**

(Source: Self-created in Google Colab)

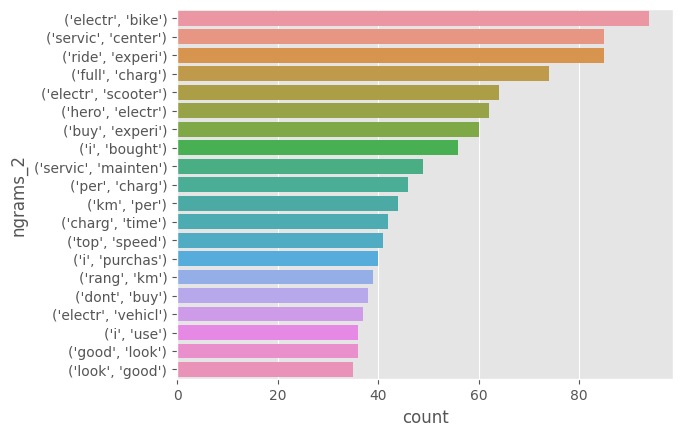
A count plot is the essential process of visualization for the data and the user point of view who is developing the analyzing system. This displays the count of the category based on the frequency of the values that are in the categorical variable (Xu, G *et al.* 2019). By the matplotlib this plot is created. This process will help to maintain the data and the notes for future purposes(Basiri *et al.* 2021). Based on the data shown in the chart graph which is based on the word and the count the understating of the main keywords which are used very frequently will be very easy and that will help to encounter the bar graph of the keywords.

****

**Figure 8: Bar plot of the data**

(Source: Self-created in Google Colab)

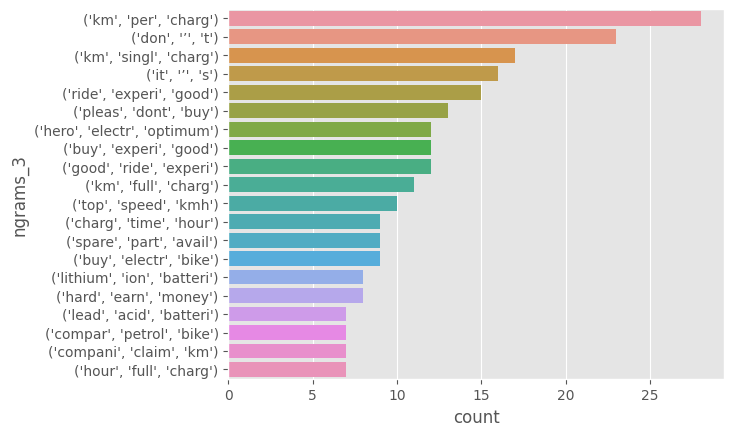
A bar plot is a very usual and common type of data visualization which used to display categorical data with bars with geographical shapes for better understanding. Matplotlib and the seaborn in Python will give the provision to display this kind of magnificent diagram based on the given data. This visualization technique is very effective for data analytics and a deep understanding of the data as well. Specifically, categorize the data in the categories then create a list according to the counts from the list of repetitive counts and create a plot based on the control size sing plot figures.

****

**Figure 9: Visual representation of BI-Diagram**

(Source: Self-created in Google Colab)

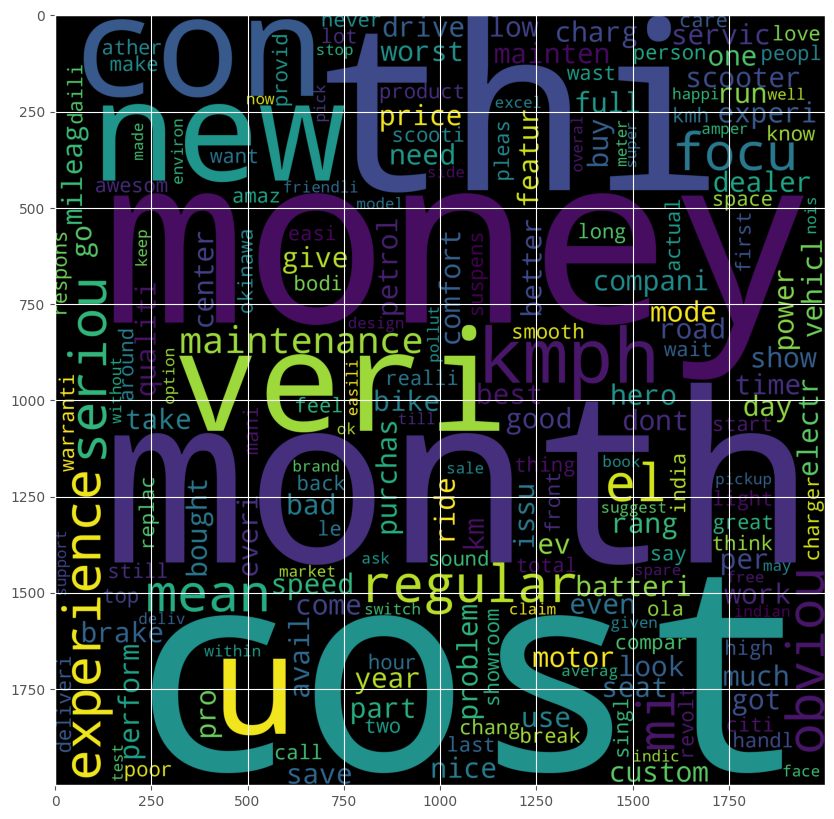
In natural language processing, this process is listed as N-gram, which are the sequences of contiguous n items from the text that is given. The two consecutive words will make a new sense which is different from a single word, good sounds positive in a monogram but for bi-gram analysis it can be so good or not good that changes the whole sense of the sentence. With the help of the NLTK package‘s punkt, this step will fulfill their tusk.

****

**Figure 10: Visual representation of TRI-Diagram**

(Source: Self-created in Google Colab)

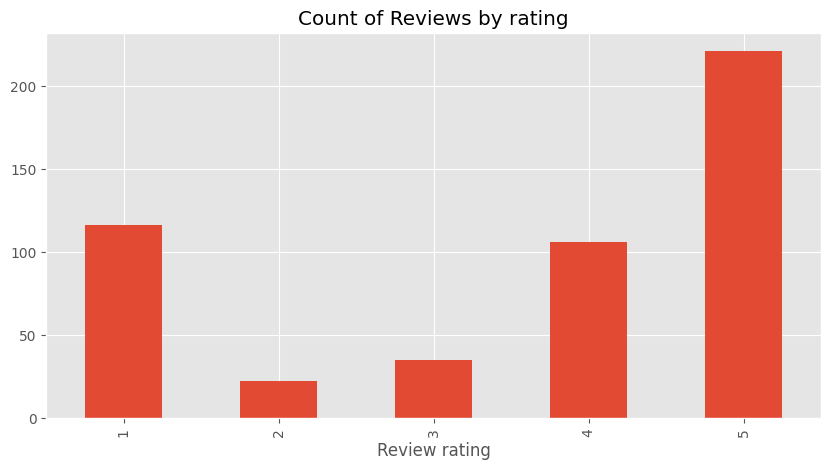
In natural language processing, this process is listed as N-gram, which are the sequences of contiguous n items from the text that is given (Yang *et al*. 2020). The three consecutive words will make a new sense which is different from a single word or two words, good sounds positive in a monogram but for tri-gram analysis it can be very unugly in nature or not so good that changes the whole sense of the sentence. With the help of the NLTK package‘s punkt, this step will fulfill their tusk. These steps are from the visualizations in the report and the next step of the analysis.

****

**Figure 11: Result of world frequency analysis**

(Source: Self-created in Google Colab)

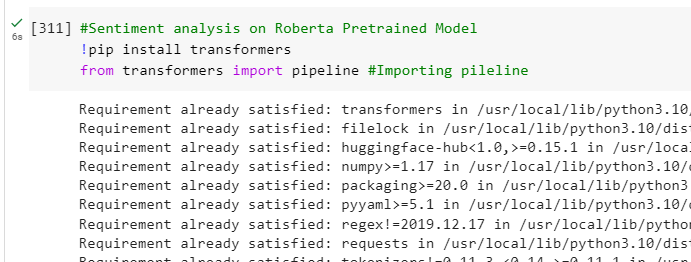
Word cloud is a process of visualization of the word frequency which is provided as the text document, this process is very popular for seeing the data frequency in a visual manner. This Cloud is a graph and the matrix form of the collective key data which are used in the dataset repetitively(Saad *et al.* 2019). These visualization techniques offer a more vivid and prominent font for a better visualizing experience. Matplot lib is used here for the displaying pot and the word cloud for the word cloud-based library. Here like an HTML front sheet, one can display thor pictures of graphs according to thie way of orientations and structural manner.

****

**Figure 12: Graph of a count of reviews and rating**

(Source: Self-created in Google Colab)

This graph shows the review and the count rating means the words that are used here more frequently. A count plot is the essential process of visualization for the data and the user point of view who is developing the analyzing system. This displays the count of the category based on the frequency of the values that are in the categorical variable. By the matplotlib, this plot is created. This process will help to maintain the data and the notes for future purposes. Based on the data shown in the chart graph which is based on the word and the count the understating of the main keywords which are used very frequently will be very easy and that will help to encounter the bar graph of the keywords.

****

**Figure 13: Implying sentiment analysis using Roberta Pre-trained**

**Model**

(Source: Self-created in Google Colab)

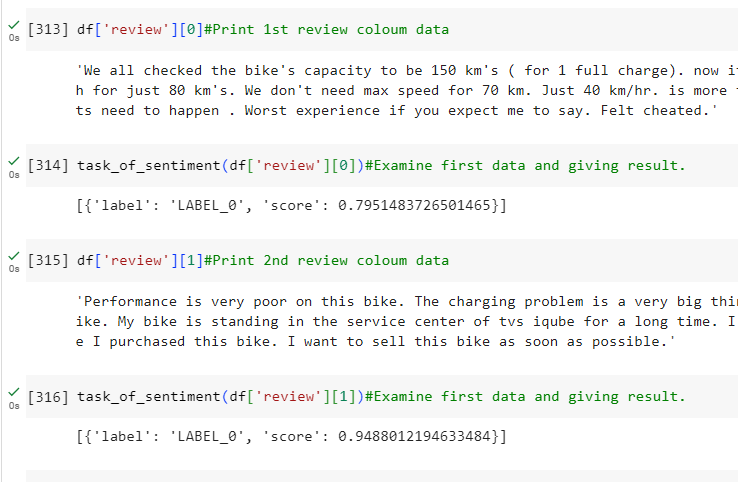
Here the chosen model is pre-trained by the RoBERTa approached alogithem. RoBERTa stands for the Robustly Optimized Bert Pretraining Approach is a very effective and efficient variant of BERT which stands for the Bidirectional Encoder Representations from the Transformers. This model is pre-trained in the NLP-based deep learning Roberta approach which gives the best result for the given data as the analysis. This was used on Facebook for the first time for AI research purposes which is why this technique is well established. This feature has large-scale penetration capabilities along with masked modeling on the language, sized batches on the dynamical method, and no sentence order prediction, data augmentation, and hyperparameter tuning.

****

**Figure 14: Train the model with the Twitter roberta data set**

(Source: Self-created in Google Colab)

The model that is used here is based on the Roberta algorithm and the model specifically used here is trained by the Twitter dataset very intensively. This model has seen the various comments reviews and sentiments from Twitter’s huge dataset (Birjali *et al.* 2021). That is why this data model has been chosen here because after being trained by the huge dataset it has worked very efficiently. We create a data pipeline and then through that the whole data of the review column of the dataset.

****

**Figure 15: Taking the first and second lines from the review to examine the review’s sentiment**

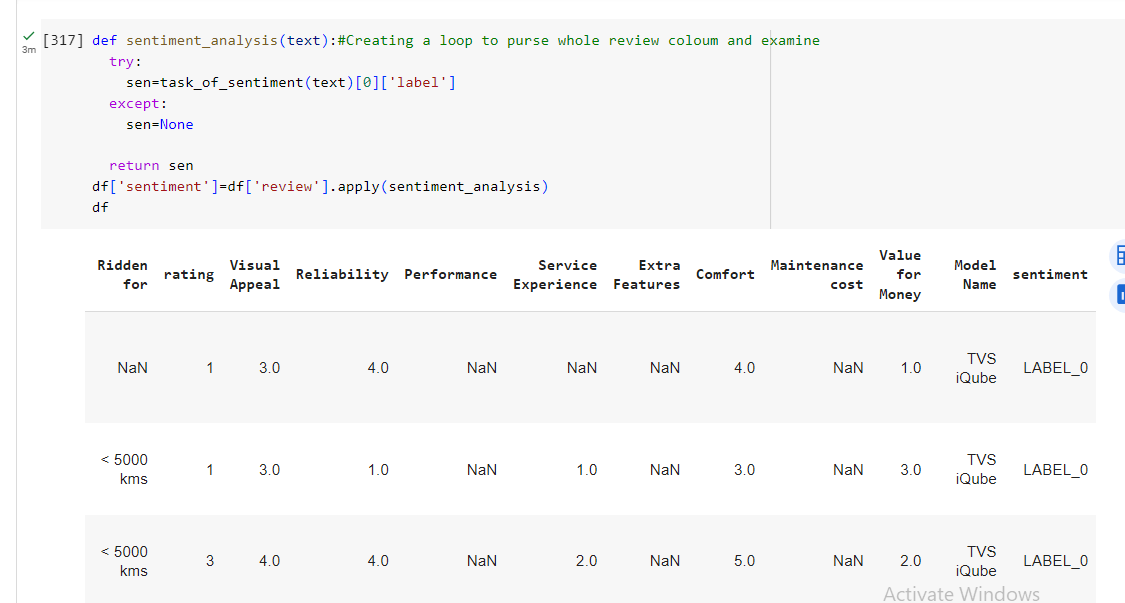
(Source: Self-created in Google Colab)

Here the test occurs before the final result, by splitting the dataset only 1st and 2nd review. After these two occurrences, this will determine that this model is working properly and with less redundancy(Wankhade *et al.* 2022). According to the data model,

LEVEL 0 > Means negative review above 0.50 in numerics.

LVELE 1 > Means neutral review nearly 0.50 in numerics.

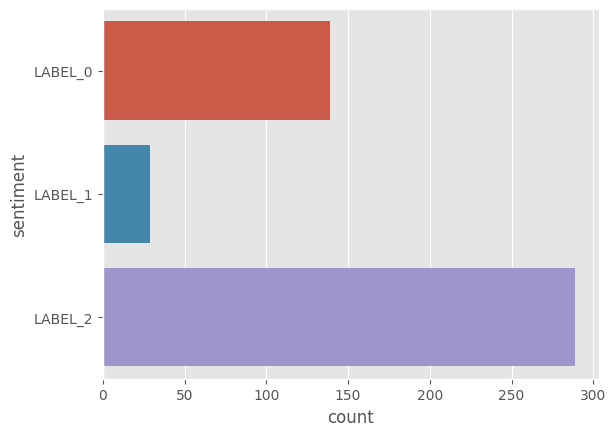
LEVEL 2 > Means positive review below 0.50 in numerics.

****

**Figure 16: Pursuing the whole data of review to predict the sentiment**

(Source: Self-created in Google Colab)

This part of the code has created a loop for inputting all the data from the mode and displays the results in a graph that is used to make predictions about the scores.

****

**Figure 17: Result of sentiment analysis on reviews**

(Source: Self-created in Google Colab)

This is the final result where the results of the analysis are very clear and descriptive, the positive ratings are very bold and far too large for the rest of them. This report has shown through the data that the EV bike is well-rated in social media and market reports.

# Conclusion

In this evolving world EVs are very sustainable and new. With the help of NLP techniques and data model training the company can predict the sentiment of all reviews. By those collective review parameters probability, they can understand the business trending and new prototype's nature and future scope.EV is a very new product in a competitive market world, to calculate the market for them this approach is very wise and efficient so that companies can understand the public emotions toward the products and the upgradations for the products which are needed for the prototype which is already in progress. By using Frequency Analysis, and Sentiment according to the analysis of the given data set of the EV bike this report determines the basic emotions and mindset towards the EV in general this is a very good amount of the large number of people’s reviews. This EV bike has more positive reviews than the neural as negative ones. This is very good for manufacturing companies. They can build their market according to the report. This report tends toward the EV bikes.

# Reference list

Wankhade, M., Rao, A.C.S. and Kulkarni, C., 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), pp.5731-5780.

Birjali, M., Kasri, M. and Beni-Hssane, A., 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, p.107134.

Yang, L., Li, Y., Wang, J. and Sherratt, R.S., 2020. Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE access*, *8*, pp.23522-23530.

Xu, G., Meng, Y., Qiu, X., Yu, Z. and Wu, X., 2019. Sentiment analysis of comment texts based on BiLSTM. *Ieee Access*, *7*, pp.51522-51532.

Basiri, M.E., Nemati, S., Abdar, M., Cambria, E. and Acharya, U.R., 2021. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems*, *115*, pp.279-294.

Abualigah, L., Alfar, H.E., Shehab, M. and Hussein, A.M.A., 2020. Sentiment analysis in healthcare: a brief review. *Recent advances in NLP: the case of Arabic language*, pp.129-141.

Li, W., Shao, W., Ji, S. and Cambria, E., 2022. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, *467*, pp.73-82.

Alonso, M.A., Vilares, D., Gómez-Rodríguez, C. and Vilares, J., 2021. Sentiment analysis for fake news detection. *Electronics*, *10*(11), p.1348.

Saad, S.E. and Yang, J., 2019. Twitter sentiment analysis based on ordinal regression. *IEEE Access*, *7*, pp.163677-163685.

Zunic, A., Corcoran, P. and Spasic, I., 2020. Sentiment analysis in health and well-being: systematic review. *JMIR medical informatics*, *8*(1), p.e16023.

Chakraborty, K., Bhattacharyya, S. and Bag, R., 2020. A survey of sentiment analysis from social media data. *IEEE Transactions on Computational Social Systems*, *7*(2), pp.450-464.

Zunic, A., Corcoran, P. and Spasic, I., 2020. Sentiment analysis in health and

well-being:systematic review. *JMIR medical informatics*, *8*(1), p.e16023.