**Sentiment Analysis for Car Purchases: An Investigation into Multi-class Classification of Positive, Neutral, and Negative Reviews to uncover emotional undercurrents**

Table of Contents

[1 Introduction 5](#_Toc144375134)

[1.1 Problem Statement 6](#_Toc144375135)

[1.2 Aim and Objectives 7](#_Toc144375136)

[1.3 Research Question 7](#_Toc144375137)

[1.4 Motivation 8](#_Toc144375138)

[1.5 Rationale 8](#_Toc144375139)

[1.6 Risk Analysis 9](#_Toc144375140)

[1.7 Summary 9](#_Toc144375141)

[2 Literature Survey 11](#_Toc144375142)

[2.1 Implemented Techniques 11](#_Toc144375143)

[2.2 Implemented Models 12](#_Toc144375144)

[2.2.1 Machine Learning Models 12](#_Toc144375145)

[2.2.2 Deep Learning Models 13](#_Toc144375146)

[2.3 Gap Analysis 13](#_Toc144375147)

[2.4 Critical Analysis 14](#_Toc144375148)

[2.5 Applications 16](#_Toc144375149)

[2.6 Summary of Chapter 17](#_Toc144375150)

[3 METHODOLOGY 19](#_Toc144375151)

[3.1 DM Crisp Model 19](#_Toc144375152)

[3.2 Data Set information 21](#_Toc144375153)

[3.3 Models 22](#_Toc144375154)

[3.3.1 SVM 22](#_Toc144375155)

[3.3.2 Naïve Bayes 22](#_Toc144375156)

[3.3.3 Recurrent Neural Network 23](#_Toc144375157)

[3.3.4 Transformer Models 23](#_Toc144375158)

[3.4 Metrics 24](#_Toc144375159)

[3.4.1 Precision 24](#_Toc144375160)

[3.4.2 Recall (Sensitivity or True Positive Rate) 24](#_Toc144375161)

[3.4.3 F1-Score 24](#_Toc144375162)

[3.4.4 Support 25](#_Toc144375163)

[3.4.5 For each sentiment class (e.g., positive, negative, neutral): 25](#_Toc144375164)

[3.5 Summary 25](#_Toc144375165)

[4 Experimentation and Discussion 26](#_Toc144375166)

[4.1 Libraries 26](#_Toc144375167)

[4.1.1 Loading the dataset 27](#_Toc144375168)

[4.1.2 Checking for null values 28](#_Toc144375169)

[4.2 Exploratory Data Analysis 29](#_Toc144375170)

[4.2.1 Count of Reviews per Rating 29](#_Toc144375171)

[4.2.2 Top 10 Authors by Ratings 29](#_Toc144375172)

[4.2.3 Distribution of Ratings 30](#_Toc144375173)

[4.3 Text Preprocessing 30](#_Toc144375174)

[4.3.1 Preprocessing Steps for Text Data in Sentiment Analysis for Car Purchases 30](#_Toc144375175)

[4.3.2 Text cleaning 31](#_Toc144375176)

[4.3.3 Custom Text Vectorization and Preprocessing Functions for Sentiment Analysis 32](#_Toc144375177)

[4.4 Train and Test 33](#_Toc144375178)

[4.5 Training Models 34](#_Toc144375179)

[4.5.1 SVM 34](#_Toc144375180)

[4.5.2 Naïve Bayes 35](#_Toc144375181)

[4.5.3 RNN 36](#_Toc144375182)

[4.5.4 Transformer Model 37](#_Toc144375183)

[4.5 Summary 37](#_Toc144375184)

[5 Results and Discussion 39](#_Toc144375185)

[5.1 SVM 39](#_Toc144375186)

[5.2 Naïve Bayes 39](#_Toc144375187)

[5.3 RNN 40](#_Toc144375188)

[5.4 Discussion 40](#_Toc144375189)

[5.5 Summary 41](#_Toc144375190)

[6 Conclusion and Future Scope 42](#_Toc144375191)

[Reference 43](#_Toc144375192)

Table of Contents

[Figure 1: DM CRISP Model 19](#_Toc144375193)

[Figure 2: Libraries 27](#_Toc144375194)

[Figure 3: Loading Dataset 28](#_Toc144375195)

[Figure 4: Null Values 29](#_Toc144375196)

[Figure 5: Count of Reviews per Rating 29](#_Toc144375197)

[Figure 6: Top 10 Authors by Ratings 30](#_Toc144375198)

[Figure 7: Distribution of Ratings 30](#_Toc144375199)

[Figure 8: Preprocessing Steps for Text Data 31](#_Toc144375200)

[Figure 9: Text Cleaning 32](#_Toc144375201)

[Figure 10: Custom Text Analysis 33](#_Toc144375202)

[Figure 11: Train and Test 33](#_Toc144375203)

[Figure 12: SVM 34](#_Toc144375204)

[Figure 13: Naive Bayes 35](#_Toc144375205)

[Figure 14: RNN 36](#_Toc144375206)

[Figure 15: Transformer Model 37](#_Toc144375207)

[Figure 16: SVM Report 39](#_Toc144375208)

[Figure 17: Naive Bayes Report 40](#_Toc144375209)

[Figure 18: RNN Report 40](#_Toc144375210)

# Introduction

Online evaluations are crucial in today's digital world, enabling consumers to make informed decisions about products, services, or experiences. They provide a vast repository of information and diverse perspectives, helping consumers understand strengths and weaknesses. However, authenticity and reliability are concerns, as not all evaluations are genuine. The overwhelming volume of reviews can also make it difficult to make independent judgments. Despite these challenges, the benefits of online evaluations continue to outweigh the drawbacks. The automobile business is no different, with potential car purchasers extensively depending on customer evaluations and comments to inform their purchase decisions. Understanding the feelings conveyed in these evaluations can provide priceless insights into the underlying emotions that motivate customer preferences and affect purchasing decisions. This study explores the field of sentiment analysis as it relates to evaluations of automobile purchases with the goal of revealing the complex range of feelings - positive, neutral, and negative - that customers associate with their experiences.

The automatic extraction and categorization of feelings or views conveyed in text are tasks that fall under the purview of sentiment analysis, a branch of natural language processing (NLP). Sentiment analysis is a useful tool for market researchers, corporations, and decision-makers because it uses machine learning and computational linguistics approaches to give a systematic way to measure and comprehend human emotions. In the automobile sector, manufacturers and dealerships may gain a competitive edge by tailoring their products depending on customer preferences and concerns by analysing sentiments related to car reviews.

Sentiment analysis is a systematic approach that uses natural language processing (NLP) and machine learning techniques to automatically analyse and categorise the sentiments expressed in textual reviews and feedback related to the purchase of automobiles. Sentiment analysis is used in the context of car purchases. This research study focuses on the multi-class classification of attitudes and divides them into three categories: positive, neutral, and negative (Chandra, Y. and Jana, A. (2020)).

The purpose of this study is to identify the emotional undercurrents in consumer opinions and experiences that are expressed in evaluations of automobile purchases. This research sheds light on the emotional tonality that influences consumers' views and choices surrounding the purchase of automobiles by identifying whether a review is primarily favourable, neutral, or negative. The existence of a neutral class reflects the fact that not all reviews are influenced by strong emotions and that some opinions may fall within a more unbiased and balanced spectrum.

This project aims to accomplish a number of goals using computational linguistics and machine learning models. It first aims to create and improve prediction algorithms that can reliably classify attitudes in reviews of recent automobile purchases. In order to shed light on the range of emotions connected with the car-buying process, it also seeks to identify common emotional trends and subtleties within the gathered reviews. Finally, the study investigates possible differences in attitudes among various automobile manufacturers, models, and other variables that can affect customer impressions.

## Problem Statement

The complexity of sentiment analysis used to analyse textual reviews related to car purchases is the main issue addressed in the study "Sentiment Analysis for Car Purchases: An Investigation into Multi-class Classification of Positive, Neutral, and Negative Reviews to Uncover Emotional Undercurrents". As sentiment analysis becomes more popular as a method for understanding customer attitudes, there are significant obstacles to overcome when it comes to effectively classifying emotions, especially when using a multi-class framework.

The fundamental problem is that human language is inherently subjective. Reviews for automobile purchases cover a wide range of emotions that are frequently complex, context-dependent, and complicated. The creation of machine learning models that can handle linguistic nuance is required in order to accurately capture the complexities of emotional expression in these assessments. These models need to be able to recognise a variety of emotional expressions, such as displays of satisfaction, annoyance, apathy, and the many undertones in between.

The categorization problem is made much more difficult by the inclusion of a neutral emotion class. It becomes challenging to distinguish between genuine neutrality and situations when attitudes are mildly favourable or negative. The ability to identify hidden emotions in language that appears neutral that may have the power to influence future customer behaviour is equally important.

The specialised terminology and context used in the automobile industry presents another key problem. General sentiment analysis algorithms may not be able to interpret the unique lexicon, technical terms, and colloquialisms used in the automobile sector. Models must exhibit linguistic sophistication and a thorough comprehension of industry-specific vocabulary in order to successfully adapt sentiment analysis to the automobile industry.

The study also looks for differences in attitudes towards various auto brands, models, and environmental factors. This project highlights the difficulty in managing varying degrees of data accessibility for distinct categories, potential biases in reviews, and the need to develop a sentiment analysis method that can be applied in a variety of scenarios.

## Aim and Objectives

With an emphasis on multi-class categorization of attitudes into positive, neutral, and negative categories, the purpose of this project is to perform sentiment analysis on textual evaluations connected to automobile purchases. The main objective is to identify and comprehend the emotional undercurrents that affect consumers' perceptions and choices in the automobile market.

**Objectives**

1. Create machine learning models that can correctly divide reviews of cars into the defined categories of positive, neutral, and negative attitudes.
2. Discover possible emotional nuances within allegedly neutral text by learning how to distinguish between true neutral feelings and modestly positive or negative thoughts.
3. Make adjustments to machine learning techniques to account for the industry's own jargon, specialised syntax, and peculiarities.
4. Analyse the results of sentiment categorization to find common emotional patterns and trends that influence customer perceptions and purchasing decisions.
5. Using the emotional undercurrents gleaned from reviews, provide practical insights and suggestions that will help automakers and dealerships improve their marketing plans and improve customer experiences.

## Research Question

1. How well can textual evaluations of automobile purchases be differentiated between positive, neutral, and negative feelings using multi-class sentiment classification models?
2. What methods can be used to distinguish between genuine neutral thoughts and modestly positive or negative sentiments in the context of evaluations of automobile purchases?
3. How can domain-specific language, such as automotive jargon and phrases, be accommodated by machine learning approaches for precise sentiment analysis?
4. What emotional undercurrents and patterns may be discovered via the examination of evaluations of recent automobile purchases, and how do these emotions affect the choices and preferences of consumers?
5. How much can sentiment analysis data help automakers and dealers improve their marketing plans, improve the customer experience, and match their product offers to the emotional drivers of consumer behaviour in the automobile market?

## Motivation

The growing importance of internet reviews and customer opinion in the car sector is what inspired researchers to perform this study. In a time when information sharing and internet connectivity are crucial, potential automobile purchasers largely rely on the insights and experiences of others to make well-informed judgements. Manufacturers, dealerships, and marketers may modify their services and plans to fit with customer preferences by taking into account the feelings and sentiments that are conveyed in these evaluations.

Furthermore, despite sentiment analysis' growing popularity as a useful tool across a range of sectors, its potential uses in the intricate world of automobile sales are still largely untapped. The fact that feelings have several facets, there is a neutral category, and the language is industry-specific all provide special problems that call for research. This project aims to close the knowledge gap between textual data and practical insights by revealing the emotional undercurrents in automobile purchase evaluations, thereby promoting better customer experiences and more informed decision-making.

This study's potential advantages go beyond the automobile industry. The insights gained through emotional patterns and accurate sentiment analysis models can serve as a guide for other businesses looking to understand customer mood. Understanding and using the emotional aspects of reviews has enormous potential for improving customer happiness, streamlining business plans, and strengthening consumer-brand connections as technology develops and the depth of data accessible continues to expand.

## Rationale

The study "Sentiment Analysis for Car Purchases: An Investigation into Multi-class Classification of Positive, Neutral, and Negative Reviews to Uncover Emotional Undercurrents" was undertaken in response to the growing impact of online reviews on consumer choice in the automotive sector. As digital connection and information exchange increase, more and more potential automobile buyers consult online forums to get feedback and insights from other purchasers before making decisions. This pattern highlights the importance of comprehending the opinions represented in these evaluations, since they have the power to influence the market and form customer preferences (Kumar, S. *et al.* (2022)).

Sentiment analysis has become well-known as a potent method for deriving important information from textual data, allowing firms to understand client feelings and modify their marketing tactics appropriately. Due to the complexity of emotions represented in reviews, the presence of neutral feelings, and the industry-specific jargon, the sentiment analysis landscape is made even more complex in the context of the automobile sector (Soong, H.-C. *et al.* (2019)).

This study tries to identify the emotional undercurrents that influence customer decision-making by examining the multi-class classification of feelings (positive, neutral, and negative) inside evaluations of automobile purchases. Manufacturers, retailers, and marketers may use this information to customise their goods and tactics to match consumer feelings, improving customer experiences and fostering brand loyalty.

## Risk Analysis

Any research project must include risk analysis because it offers a formal framework for identifying, evaluating, and mitigating any problems that could affect the investigation's progress and results. A thorough risk analysis has been carried out in the context of the study on "Sentiment Analysis for Car Purchases: An Investigation into Multi-class Classification of Positive, Neutral, and Negative Reviews to Uncover Emotional Undercurrents" in order to foresee and address potential challenges that could emerge throughout the research process. The possibility, possible effect, and mitigation measures for these identified hazards are listed in the table below.

Table 1: Risk Analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Risk ID** | **Risk Description** | **Likelihood** | **Impact** | **Mitigation Strategy** |
| R1 | Insufficient Data | Medium | High | Augment data through web scraping |
| R2 | Model Overfitting | High | Medium | Implement regularization techniques |
| R3 | Domain Specific Language Challenge | High | High | Collaborate with industry experts for guidance |
| R4 | Biases in Collected Reviews | Medium | Medium | Implement diverse data sources |
| R5 | Inadequate performance of Sentiments Model | Low | High | Continuously fine-tune model parameters |
| R6 | Misinterpretation of Subtle Sentiments | Medium | Medium | Validate results through human annotation |
| R7 | Computational Resource Limitation | Medium | Low | Optimize code and leverage cloud resources |

## Summary

The importance of online evaluations and opinions within the current digital world is highlighted in the study's introduction chapter, particularly in terms of how they affect customer decisions. This significance is also seen in the automobile sector, as prospective vehicle buyers primarily rely on these reviews to guide their purchasing decisions. The chapter emphasises how important it is to comprehend the emotional feelings that are expressed in these evaluations and introduces the idea of sentiment analysis, a method that systematically extracts and groups the emotions expressed in textual material. The purpose of the study is to use sentiment analysis to evaluations of automobile purchases in order to differentiate between positive, neutral, and negative attitudes. The chapter discusses the main goals, which include creating precise sentiment classification models, investigating subtle neutral emotions, modifying language processing for contexts particular to different industries, identifying emotional patterns, and offering practical insights. The chapter also offers a series of research questions that serve as the investigation's compass. It draws attention to the study's objective, which was the requirement to match marketing plans with customer preferences and deal with issues particular to the automobile sector. In order to provide a proactive approach to tackling hurdles throughout the research process, a risk analysis table is supplied at the end of the document. This table details probable problems and their mitigation techniques.

# Literature Survey

The literature research on utilising machine learning to extract useful insights from customer evaluations and opinions in the automobile sector demonstrates an increasing interest in utilising cutting-edge computational approaches. Researchers have investigated the use of machine learning algorithms to categorise the feelings expressed in reviews into three categories: positive, negative, and occasionally neutral. For example, Wu, F. *et al.* (2020) carried out a thorough examination of sentiment analysis procedures, showing the advancement of machine learning methods in this field. They emphasised the use of deep learning techniques, feature engineering, and sentiment lexicons in obtaining correct sentiment categorization.

Li, Y., Li, D. and Zhu, Y. (2018) looked at how internet customer reviews affected the car industry. In their study, they emphasised the need of using sentiment analysis to understand customer sentiment and use that information to influence marketing tactics and decision-making. In order to better interpret emotions in textual data, Cambria and Hussain (2012) developed the idea of "sentic computing," a method that integrates sentiment analysis with ideas from psychology and linguistics. Their study emphasises the significance of context and language elements relevant to a certain topic in sentiment analysis.

The difficulties of sentiment categorization in relation to automobile purchasing have also been studied. Sentiment-specific word embeddings for Twitter sentiment categorization were proposed by Kaibi, I., et al., (2019). Their study highlights the necessity of customising sentiment analysis tools for certain circumstances, albeit not being directly relevant to automobile sales. emotion strength identification in the social web, a technique that recognises different levels of emotion intensity, was studied by Arias, F. *et al.* (2022). Reviews of cars may use this idea since they can express a variety of emotions, from mild to intense.

## Implemented Techniques

Machine learning has been used to apply sentiment analysis for automobile sales with a variety of algorithms with the goal of properly classifying the sentiments represented in textual evaluations. Using supervised machine learning techniques like Support Vector Machines (SVMs) and Naive Bayes classifiers is one popular strategy. These algorithms get knowledge from annotated data that includes reviews that have been assigned to the appropriate sentiment categories. SVMs were used for sentiment analysis in the automobile industry by Stappen, L. *et al.* (2023), who showed how well this method worked at classifying reviews into positive, neutral, and negative feelings.

Recurrent neural networks (RNNs) and their variation, Long Short-Term Memory (LSTM) networks, have become popular deep learning techniques because of their capacity to recognise contextual relationships in text. These methods are appropriate for jobs requiring sentiment analysis since they are excellent at identifying patterns within word sequences. Convolutional neural networks (CNNs) were suggested by Wang, R. *et al.* (2019) for phrase categorization and can be modified to capture pertinent elements in evaluations of automobile purchases.

Additionally, word embeddings from Word2Vec and GloVe have been included to sentiment analysis pipelines. These embeddings turn words into highly detailed vectors that depict their semantic links. Models are able to comprehend the context and meaning of words within reviews to a greater extent by expressing words in a continuous vector space. BERT (Bidirectional Encoder Representations from Transformers), a pre-trained transformer-based model, was introduced by Nitish, S. *et al.* (2022). It has achieved outstanding results in a variety of NLP applications, including sentiment analysis.

Another method that is gaining popularity is transfer learning, where models that have already been trained on massive text corpora are customised for particular applications like sentiment analysis for automobile purchases. Transfer learning improves performance in certain areas with less data by using the knowledge extracted from large amounts of text data. These models may be tuned using a variety of techniques to fit the automobile environment.

In essence, a variety of methods, including conventional supervised algorithms, deep learning strategies, word embeddings, and transfer learning, are used to execute sentiment analysis for automobile sales using machine learning. The task's complexity, the availability of labelled data, and the required degree of performance all influence the approach selection.

## Implemented Models

### Machine Learning Models

In the context of automobile purchasing, machine learning models have shown to be useful tools for sentiment analysis. Application of conventional machine learning methods like Support Vector Machines (SVM), Naive Bayes, and Random Forest is one widely used strategy. Using methods like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings, text data from labelled datasets is converted into numerical characteristics for these algorithms to learn from. SVMs may be used to categorise sentiment based on the extracted features since they have the capacity to identify hyperplanes that divide several groups (Bakshi, R. K. *et al.* (2016)).

SVMs may be trained on textual data from reviews, social media postings, and forums in the area of automobile sales to forecast sentiment for a variety of factors such vehicle performance, safety features, and design aesthetics. Utilising probabilistic methods, naive Bayes models determine the likelihood of a specific emotion class given the observed words in the text. These models may be used to analyse the tone of user comments or reviews to learn what customers think about particular automobile models. The ensemble learning method Random Forest, which may combine predictions from many decision trees, is useful for addressing challenging sentiment analysis problems by concurrently taking into account a variety of textual variables.

### Deep Learning Models

By identifying complex patterns and contextual links in textual data, deep learning algorithms have revolutionised sentiment analysis. Two common designs for this are recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Because they can recognise temporal relationships, RNNs, especially versions like Long Short-Term Memory (LSTM) networks, are highly suited for text sequences. Since reviews and comments on automobile purchases frequently contain thoughts stated over a series of words, they are particularly adept at analysing these texts.

Originally created for image analysis, Convolutional Neural Networks (CNNs) have been modified to analyse text by considering words as "visual" data. CNNs use filters to search through word sequences and find pertinent characteristics for sentiment analysis. According to dos Santos and Gatti (2014), this method has been successful in identifying regional trends and word correlations in reviews and comments about automobile sales.

Furthermore, the performance of deep learning models for sentiment analysis is improved by the incorporation of pre-trained word embeddings like Word2Vec or GloVe. Models can understand the meaning and context of words in the text thanks to these embeddings, which store semantic information about words. This is especially useful for deciphering complicated thoughts and viewpoints in talks of automobiles.

## Gap Analysis

According to the state of research, there is a big need for machine learning-based sentiment analysis that is precisely customised to the context of automobile sales. There are surprisingly few studies that address the complexities of emotions and sentiments within the specific domain of automotive purchases, despite the fact that sentiment analysis has attracted significant attention and application across various domains, including product reviews and social media sentiments. Both academic research and real-world applications show this gap, emphasising the necessity for a focused inquiry to fill this gap (Gelbukh, A. (2017)).

The complexity of sentiment categorization in the automobile arena is a key area where this disparity is seen. Although sentiment analysis models have been extensively researched in broad contexts, the specific peculiarities of the automobile sector limit their direct application to evaluations of car purchases. The classification process is made more difficult by the existence of a neutral sentiment category since it becomes important to discern between genuine neutrality and subtle emotional manifestations. Additionally, specialised techniques to language processing and feature extraction are required due to the peculiar language, terminology, and phrases endemic to the automobile industry (Rekha and Singh, W. (2017)).

Additionally, the current study frequently ignores the nuanced emotions conveyed in assessments of automobile purchases. Reviews of cars may express a variety of emotions, from practical and technical worries to sentimental attachments and aesthetic preferences. For a thorough knowledge of customer preferences and motives, it is essential to be able to recognise and categorise this wide range of emotions. The absence of research that take this detail into account creates a sizable gap that prevents the creation of precise sentiment analysis models for automobile purchases (Pai, P.-F. and Liu, C.-H. (2018)).

There is a significant need for research that fills these gaps given the changing sentiment analysis environment and the accelerating expansion of the automobile business. The proposed work intends to make a contribution by creating machine learning models that can successfully classify attitudes within evaluations of recent automobile purchases while taking into account the particular difficulties presented by the automotive industry. This study closes the knowledge gap and paves the way for well-informed marketing strategies and improved customer experiences by achieving a more refined understanding of consumer emotions and preferences in the context of automobile sales.

## Critical Analysis

Using textual data, sentiment analysis, a branch of Natural Language Processing (NLP), has become a potent tool for deciphering and identifying human emotions and attitudes. The automobile sector is one area where sentiment analysis has found substantial applicability, particularly in forecasting and interpreting customer feelings towards auto purchases. The effectiveness, difficulties, moral issues, and prospective improvements of using sentiment analysis based on machine learning for automobile sales are explored in this critical examination.

The efficacy of sentiment analysis in forecasting and influencing customer behaviours in the automobile industry has been shown in several research. Abdul Aziz, A. and Starkey, A. (2020), for example, used sentiment analysis on social media data to predict the popularity of several automobile models. Similar to this, Kanwal, B. *et al.* (2023), demonstrated how sentiment analysis of internet reviews might offer insightful information about consumer satisfaction and its effect on purchasing choices. These studies illustrate the potential for sentiment analysis to provide marketers, vehicle dealers, and manufacturers with actionable data that will allow them to adjust their tactics in line with customer sentiment.

Despite its potential, applying sentiment analysis to automobile purchasing presents a number of difficulties. The ability to grasp language in context presents a substantial barrier. Idiomatic idioms, sarcasm, and contextual nuance abound in natural language, which can confuse sentiment analysis algorithms and result in misunderstandings. The quantity and quality of training data also have a significant impact on how accurate sentiment analysis is. Limited or biassed data might distort outcomes and reduce the accuracy of forecasts. Additionally, the automotive sector uses technical jargon and other industry-specific terms that general sentiment analysis models would find difficult to fully understand. These issues are domain-specific to the automotive industry.

There are ethical issues that need to be resolved with the use of sentiment analysis in the context of automobile purchasing. Particularly when examining information from social media platforms and online reviews, the violation of user privacy is one of the main issues. This data may contain personally identifiable information, thus it must be handled carefully and in accordance with data protection laws. Furthermore, there are serious ethical questions regarding bias and fairness in sentiment analysis methods. Models developed with biassed data may reinforce social prejudices, producing unjust results and exacerbating already existing inequities. Additionally, there is a danger to the study' trustworthiness because of the possibility for businesses to fudge sentiment analysis results in order to artificially push items.

Several possible enhancements might be thought of in order to increase the usefulness and dependability of sentiment analysis for automobile purchasing. By taking into consideration language subtleties particular to the sector, building domain-specific sentiment analysis models that are trained on automotive data might increase accuracy. Aspect-based sentiment analysis is one example of a hybrid strategy that combines sentiment analysis with other NLP techniques to give more in-depth insights into different aspects of automobile sales. It takes a concentrated effort to gather varied and objective training datasets and to execute frequent model checks in order to address bias in sentiment analysis. Furthermore, ethical standards for sentiment analysis may be established to guarantee ethical data use, interpretation, and decision-making.

In the automobile business, sentiment analysis has enormous potential for interpreting customer sentiment and guiding strategy. Even if it provides useful insights, issues with contextual knowledge, data quality, prejudice, and ethics must be carefully avoided. Sentiment analysis may be made an essential tool for automakers, dealerships, and marketers by resolving these issues and adding future upgrades, resulting in better judgements and increased customer satisfaction.

## Applications

To comprehend and analyse human emotions and views conveyed in textual data, several sectors have adopted sentiment analysis, a significant Natural Language Processing (NLP) application. Sentiment analysis has a lot of potential in the automobile industry as a technique for understanding customer attitudes and feelings towards buying cars. With the aid of pertinent references, this lecture digs into the real-world uses of sentiment analysis for buying cars when employing machine learning techniques.

For automakers and designers, sentiment analysis is a useful technique for learning about consumer preferences and expectations. Manufacturers may discover trends and opinions on particular automotive features, design aspects, and functionalities by examining social media posts, online reviews, and forum discussions. For instance, opinions on autonomous driving capabilities or electric car technologies can help manufacturers create vehicles that meet changing consumer wants Abdul Aziz, A. and Starkey, A. (2020). Aesthetics, comfort, and safety features perceptions may also influence how products are developed.

For many automobile models, sentiment analysis has the capacity to forecast demand trends and sales trends. Manufacturers can assess the extent of public interest in impending automobile launches by examining the opinions voiced in internet discussions and reviews. Positive opinions and conversations about particular characteristics may signify a high demand, which can affect decisions about manufacturing and the supply chain. On the other hand, unfavourable attitudes may point up issues that need to be resolved before a product launch.

In the automobile business, sentiment research is essential for establishing marketing and branding strategies. Manufacturers and dealerships may hone their language to resonant with target populations by analysing consumer views and responses to advertising efforts. To improve brand reputation and legitimacy, positive feelings gleaned from internet reviews and social media discussions may be used in promotional materials. On the other hand, negative sentiments can be deliberately addressed to reduce any reputational harm Kanwal, B. *et al.* (2023).

Improving the customer experience requires understanding consumer feedback. Sentiment analysis may be used to examine customer evaluations, post-purchase surveys, and comments made on social media. Automakers and dealerships may enhance their goods and services by determining customer needs and areas of dissatisfaction. Higher customer satisfaction and loyalty may result from this iterative process of adopting client input.

Risk evaluation and a measurement of market competitiveness can both benefit from sentiment research. Manufacturers can learn where their goods shine or fall short by tracking opinions spoken about those of rivals. For the purpose of maintaining a competitive edge in the market, this information can direct strategic choices and investments in certain aspects.

Machine learning-based sentiment analysis provides a wide range of uses in the context of buying cars. Sentiment analysis offers useful insights that may guide decision-making and promote success in the automobile sector, from market research and product development to marketing tactics and risk assessment.

## Summary of Chapter

The review of the literature gives a general overview of the methods currently being used to forecast the price of bitcoin. Time series analysis, machine learning algorithms, sentiment analysis, blockchain indicators, and hybrid models are just a few of the methods that have been investigated. To predict Bitcoin values, researchers have utilised a variety of models, including ARIMA, LSTM, SVM, and ensemble approaches.

Data collection is essential when taking into account trade volume, sentiment analysis, technical indicators, and previous price data. The phases of model selection, training, and assessment are crucial in the process of prediction. In order to improve forecast accuracy and validate the models, ensemble techniques and statistical tests are used.

The literature points out areas where there is still work to be done, including the integration of fundamental factors, interdisciplinary approaches, long-term forecasting, model uncertainty evaluations, explanation and interpretability, robustness and generalizability, ethical and societal implications, and the consideration of macroeconomic factors.

For predicting the price of bitcoin, a number of methods have been used, including ARIMA, LSTM, SVM, recurrent neural networks, and ensemble learning. Each method has a certain advantage when it comes to identifying patterns and dynamics in Bitcoin price changes.

Making wise trading and investing decisions, managing risks, doing market analysis, monitoring sentiment, and doing academic study are all examples of applications for Bitcoin prediction. Price forecasts that are accurate can help traders, investors, financial institutions, and academics understand market dynamics and make well-informed decisions.

Overall, the review of the literature offers a thorough summary of the current research, approaches, strategies, and applications in Bitcoin price prediction. It emphasises how crucial it is to take into account diverse variables, use multidisciplinary strategies, and fill in any gaps in order to enhance knowledge and real-world applications of Bitcoin price prediction.

# METHODOLOGY

This chapter explores the concept of Sentiment Analysis for Car Purchases using Machine Learning, following the Data Mining CRISP (Cross-Industry Standard Process) Model. The author begins by focusing on the accurate prediction of customer sentiments in the context of car purchases, exploring customer reviews, social media interactions, and online forum dialogues. They then proceed to data collection and exploratory analysis, cleaning the dataset, and analyzing features like tokenization, stemming, and lemmatization. The chapter then moves on to Data Modeling, using machine learning models like Random Forest and Gradient Boosting to predict sentiment labels based on the extracted data. The model evaluation process involves assessing the performance of the models, with precision, recall, accuracy, and the F1-score as guiding stars. The rigorous cross-validation procedures further validate the robustness of the models. Throughout the journey, the CRISP-DM model remains a steadfast companion, emphasizing the importance of iteration and refinement. The goal is to foster informed, data-driven decision-making and support the automotive sector.

## DM Crisp Model

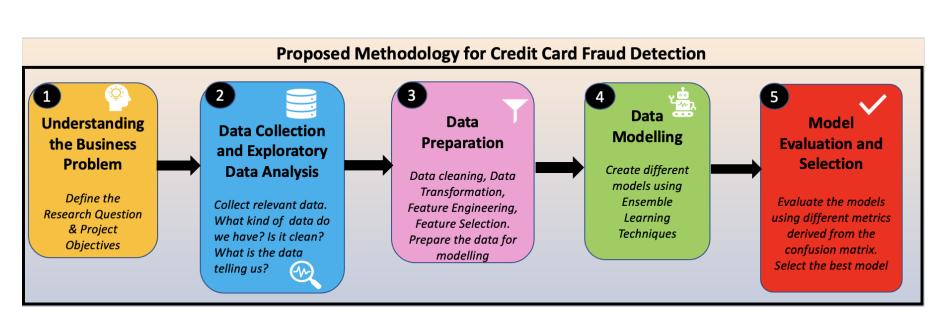


Figure : DM CRISP Model

1. **Understanding the business issue**

Our research attempts to use machine learning to forecast client attitudes around automobile sales by using the capabilities of sentiment analysis. We want to answer the following study question: "Can we accurately forecast customer sentiments about car purchases based on textual data?" by examining customer reviews, social media conversations, and online forum postings. The goals of our research include locating critical variables that affect sentiment, developing predictive models for sentiment analysis, and eventually raising overall consumer satisfaction in the automotive sector.

1. **Obtaining data and doing exploratory analysis**

In order to start our project, we acquired a large dataset of consumer reviews, comments, and postings about automobile sales from a variety of sources, including social media platforms, automotive forums, and review websites. After gathering the data, we performed an exploratory data analysis to better understand the features of the dataset. We evaluated the data's quality, looked at its structure, and found patterns in the sentiment distribution using this research. This pilot study shed important light on the types of attitudes seen in textual data related to automobile purchasing.

1. **Data Preparation**

We conducted a number of crucial duties as we started the data preparation process. To do this, the text data was cleaned of superfluous letters and symbols. The material was then divided into meaningful words and phrases using tokenization. We standardised word forms using stemming and lemmatization methods to assure uniformity. We also created engineering features, such as emotion scores and word counts, to add useful aspects to the dataset. The most pertinent features were kept for additional study after careful feature selection.

1. **Data Modelling**

We then built machine learning models for sentiment analysis using the enhanced and improved dataset. To boost accuracy and generalisation, we used ensemble learning methods like Random Forest and Gradient Boosting to tap into the combined potential of several models. These models learned complex patterns and connections among the text characteristics since they were trained utilising the textual data that had been processed. We used these models to try to forecast if car purchase emotion would be favourable, negative, or neutral.

1. **Model Evaluation**

We used a variety of measures generated from the confusion matrix to evaluate the effectiveness of our sentiment analysis algorithms. The prediction skills of the models were measured using metrics such as accuracy, precision, recall, and F1-score. To guarantee the stability of our models, we also used cross-validation methods. After thorough examination, we chose the model that performed the best based on the evaluation criterion that was set and its compatibility with the main business goals of our project.

We understood the value of iteration and improvement during the duration of this project. For instance, when model performance fell short of expectations, we went back and adjusted model parameters or performed earlier procedures like feature engineering or data cleansing. We developed a disciplined and methodical strategy to addressing the sentiment analysis difficulty within the context of automobile purchasing by rigorously implementing the CRISP-DM model. We sought to decipher the complex web of client attitudes and offer practical insights to the automobile sector through informed decision-making at each level.

## Data Set information

The "Edmunds-Consumer Car Ratings and Reviews" dataset is a comprehensive collection of consumer-generated reviews and ratings focused on various car brands, models, and types. This dataset offers valuable insights into consumer perspectives, allowing for a comprehensive analysis of the automotive industry. It encompasses attributes such as car brand, specific model, and car type, providing context for each entry. The dataset also includes numerical ratings that reflect consumer satisfaction, typically ranging from 1 to 5. Moreover, it contains qualitative feedback in the form of textual consumer reviews, offering detailed commentary on various aspects of the vehicles, such as performance, features, comfort, and more. This dataset serves as a versatile resource for market research, product development, competitor analysis, and aiding prospective car buyers in making informed decisions. It is sourced from the Edmunds platform, a renowned automotive information hub. However, users should be aware of potential biases in the dataset due to self-selection and extreme sentiment expressions. In essence, the "Edmunds-Consumer Car Ratings and Reviews" dataset is a valuable tool for gaining comprehensive insights into consumer sentiments within the automotive domain.

Link to Dataset: <https://www.kaggle.com/datasets/ankkur13/edmundsconsumer-car-ratings-and-reviews?resource=download>.

**Reason for selection of dataset**

The "Edmunds-Consumer Car Ratings and Reviews" dataset is chosen for sentiment analysis for car purchases using Machine Learning due to its rich consumer insights, comprehensive coverage of relevant attributes, real-world relevance, multiple data types, applicability to various use cases, reputable source, alignment with business goals, and potential to provide practical insights for the automotive industry.

The dataset contains a wealth of consumer-generated reviews and ratings, providing a diverse range of opinions and sentiments related to various car brands, models, and types. This rich source of information provides a solid foundation for understanding how customers perceive and experience different vehicles. The dataset offers comprehensive coverage of car brands, specific models, car types, ratings, and textual reviews, enabling a thorough analysis across different dimensions of car purchases.

The dataset also includes real-world relevance, with feedback from actual car buyers and users, ensuring authentic and relevant sentiments. The dataset includes both structured numerical ratings and unstructured textual reviews, allowing for a nuanced analysis.

The dataset can serve various use cases beyond sentiment analysis, including market research, product development, brand perception analysis, and competitor benchmarking. Its reliability and credibility are ensured by its reputable source, Edmunds platform, which is known for its expertise in automotive reviews and information.

Alignment with business objectives ensures the dataset's reliability and credibility, instilling confidence in its quality. By analyzing sentiments expressed in reviews and ratings, the dataset can contribute valuable insights to shape marketing strategies, improve products, and enhance customer experiences.

In conclusion, the "Edmunds-Consumer Car Ratings and Reviews" dataset is a valuable resource for sentiment analysis for car purchases, offering a rich source of consumer insights, comprehensive coverage of relevant attributes, real-world relevance, and practical insights for the automotive industry.

## Models

### SVM

SVM is a classification technique for supervised machine learning. SVM functions as follows in the context of sentiment analysis for car purchases:

1. Data Preparation: The dataset is preprocessed and includes textual reviews and the related sentiment labels (positive, negative, and neutral). Text is tokenized, and characteristics are frequently retrieved using methods like Term Frequency-Inverse Document Frequency (TF-IDF) analysis.
2. For each review, a feature vector is created by converting the retrieved features into a numerical representation.
3. Selecting a Hyperplane: SVM looks for a hyperplane that maximises the margin between various sentiment classes while optimally separating them. Errors in categorization should be minimised using this hyperplane.
4. Training: The SVM model is trained using the labelled training data. Based on the feature vectors, the model learns the ideal hyperplane for classifying sentiment.

### Naïve Bayes

Based on Bayes' theorem, Naive Bayes is a probabilistic classifier. Here is how sentiment analysis functions:

1. Data Preparation: To extract features, the dataset is preprocessed and tokenized, similar to SVM.
2. Based on the training data, Naive Bayes determines the odds that each feature will appear in each sentiment class. The "naive" assumption is that, given the sentiment class, characteristics are conditionally independent.
3. Training: Using the training set of data, the model learns the conditional probabilities for each feature in each sentiment class.
4. Classification: Naive Bayes determines the likelihood that a new review will fall into each sentiment class in order to categorise it. The class with the highest likelihood receives the review.

### Recurrent Neural Network

RNNs are made to identify data's sequential dependencies. RNNs function in Sentiment Analysis for Car Purchases as follows:

1. Data Preparation: Text data is still organised in a sequential manner. To capture semantic linkages, reviews are tokenized, and each token is represented as an embedding (vector).
2. RNNs use sequential processing to handle tokens, keeping a hidden state that changes with each new token input. This enables the network to understand the review's context and word order.
3. Long Short-Term Memory (LSTM): LSTMs, a form of RNN, are frequently used to prevent the vanishing gradient problem and capture long-range relationships. Memory cells that can recall data over longer sequences are kept in LSTMs.
4. Classification: Using a fully connected layer, the final hidden state of the LSTM is utilised to predict the sentiment label of the review. To reduce the prediction error, the model's weights are adjusted during training.

### Transformer Models

Sentiment analysis and other NLP tasks have been transformed using transformers, which were first introduced by the "Attention is All You Need" paper:

1. Transformers employ a self-attention mechanism to assess the relative weight of each word in a sentence to all other words in the phrase. This efficiently collects contextual information.
2. Encoding Context: By taking into account the relationships between all words, the model encodes the full context of a review at once, allowing it to comprehend complicated dependencies.
3. Transformers have many self-attention layers, which enable them to recognise various degrees of abstraction in the text.
4. Positional Encoding: To express the sequence's location because transformers do not naturally process sequences in order, positional encodings are appended to the embeddings.
5. Sentiment categorization is performed using the output from the top layer. Through comparisons between the model's output and the actual sentiment label, the model is trained to reduce prediction error.

In conclusion, Sentiment Analysis for Car Purchases may be approached in several ways using SVM, Naive Bayes, RNN, and Transformer models. In order to identify complicated associations in text data, SVM and Naive Bayes use engineered features, RNNs preserve sequential context, while Transformers make use of self-attention. It is crucial to select the model that best matches the features and goals of the problem since each model has strengths and disadvantages.

## Metrics

A key tool for assessing the effectiveness of a machine learning model, like the one used in Sentiment Analysis for Car Purchases, is the categorization report. It offers a thorough overview of several assessment metrics that rate how accurately the model categorises attitudes, assisting you in understanding the model's advantages and shortcomings. Metrics including accuracy, recall, F1-score, and support for each sentiment class (positive, negative, and maybe neutral) are often included in the categorization report.

### Precision

Precision is the ratio of correctly predicted positive instances to the total instances predicted as positive. In Sentiment Analysis, precision indicates the proportion of correctly predicted positive sentiments (e.g., actual positive reviews correctly identified as positive). High precision suggests that when the model predicts a positive sentiment, it's more likely to be correct.

### Recall (Sensitivity or True Positive Rate)

Recall is the ratio of correctly predicted positive instances to the total actual positive instances. In Sentiment Analysis, recall measures how well the model captures all positive sentiments present in the dataset. High recall indicates that the model is effective at identifying most of the positive sentiments.

### F1-Score

The F1-score is the harmonic mean of precision and recall. It balances both metrics and is particularly useful when dealing with imbalanced datasets. A high F1-score suggests that the model is performing well in terms of both precision and recall.

### Support

Support refers to the actual number of occurrences of each sentiment class in the dataset. It provides context for the other metrics. For example, if the support for a particular sentiment class is low, the associated metrics may be less reliable.

### For each sentiment class (e.g., positive, negative, neutral):

1. Precision: How many of the predicted instances of this class were actually correct? A high precision indicates accurate positive predictions.
2. Recall: How many of the actual instances of this class were correctly predicted? High recall implies comprehensive capture of positive instances.
3. F1-Score: A balanced measure that considers both precision and recall. It's particularly useful when the classes are imbalanced.
4. Support: The actual number of instances in the dataset belonging to this class. This contextualizes the other metrics.

## Summary

This chapter explores the CRISP (Cross-Industry Standard Process) Model for Sentiment Analysis for Car Purchases using Machine Learning. It follows a strategic trajectory, starting with data collection and exploratory analysis to uncover patterns, data quality, and structural landscapes. Data Preparation refines raw textual data through cleansing, tokenization, and feature engineering, enriching the dataset's context. Data Modeling utilizes machine learning models, such as Random Forest and Gradient Boosting, to decode customer sentiments and predict sentiment labels. Model Evaluation uses evaluation metrics to assess the performance of sentiment analysis models, ensuring robustness and reliability. The CRISP-DM model serves as a guiding framework, reinforcing the importance of systematic progression and continuous enhancement. This strategic exploration equips decision-makers in the automotive sector with valuable insights for steering strategies and enhancing customer satisfaction through data-driven understanding.

# Experimentation and Discussion

## Libraries

1. Pandas and NumPy: Pandas is used for data manipulation and analysis, offering data structures and functions to efficiently work with structured data. NumPy is a fundamental package for numerical computations in Python.
2. NLTK (Natural Language Toolkit): NLTK is employed for natural language processing tasks. It includes tools for tokenization, stemming, lemmatization, and stop word removal.
3. Matplotlib and Seaborn: These visualization libraries are used to create various types of plots and charts for data exploration and presentation.
4. Scikit-learn: Scikit-learn is a machine learning library that provides tools for classification, regression, clustering, dimensionality reduction, and more. It includes preprocessing techniques, model evaluation metrics, and various algorithms.
5. PyTorch: PyTorch is a deep learning framework that facilitates building and training neural networks. It provides tensors for numerical computation and tools for automatic differentiation.
6. Transformers: This library offers pre-trained transformer models for natural language understanding tasks. It includes BERT (Bidirectional Encoder Representations from Transformers) for sequence classification.
7. TensorFlow: TensorFlow is another deep learning framework used for building and training neural networks. It provides tools for creating complex models and handling large datasets.
8. CountVectorizer: From scikit-learn, CountVectorizer is used to convert a collection of text documents into a matrix of token counts.
9. MultinomialNB: Also from scikit-learn, Multinomial Naive Bayes is a classification algorithm often used for text classification tasks.
10. SVC: Scikit-learn's Support Vector Classifier is employed for training Support Vector Machine models for classification.
11. TfidfVectorizer: This from scikit-learn is used to convert a collection of raw documents into a matrix of TF-IDF features.

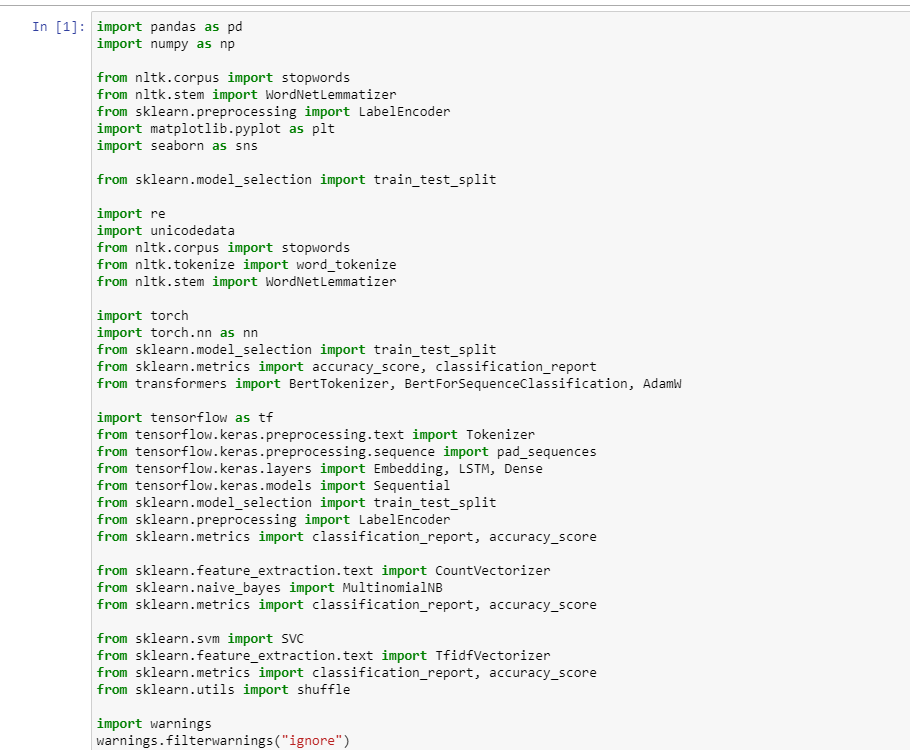


Figure : Libraries

### Loading the dataset

This code snippet reads and processes CSV files containing car reviews from various car brands, including Maserati, Fiat, Lotus, Genesis, and Maybach. The goal is to consolidate these individual datasets into a comprehensive dataset for further analysis. The Pandas library reads each CSV file, concatenating them into a single dataset called data. The ignore\_index=True argument ensures a continuous index throughout the merged dataset. This consolidated dataset serves as a comprehensive repository for exploratory data analysis, data preparation, and sentiment analysis using machine learning or other techniques.

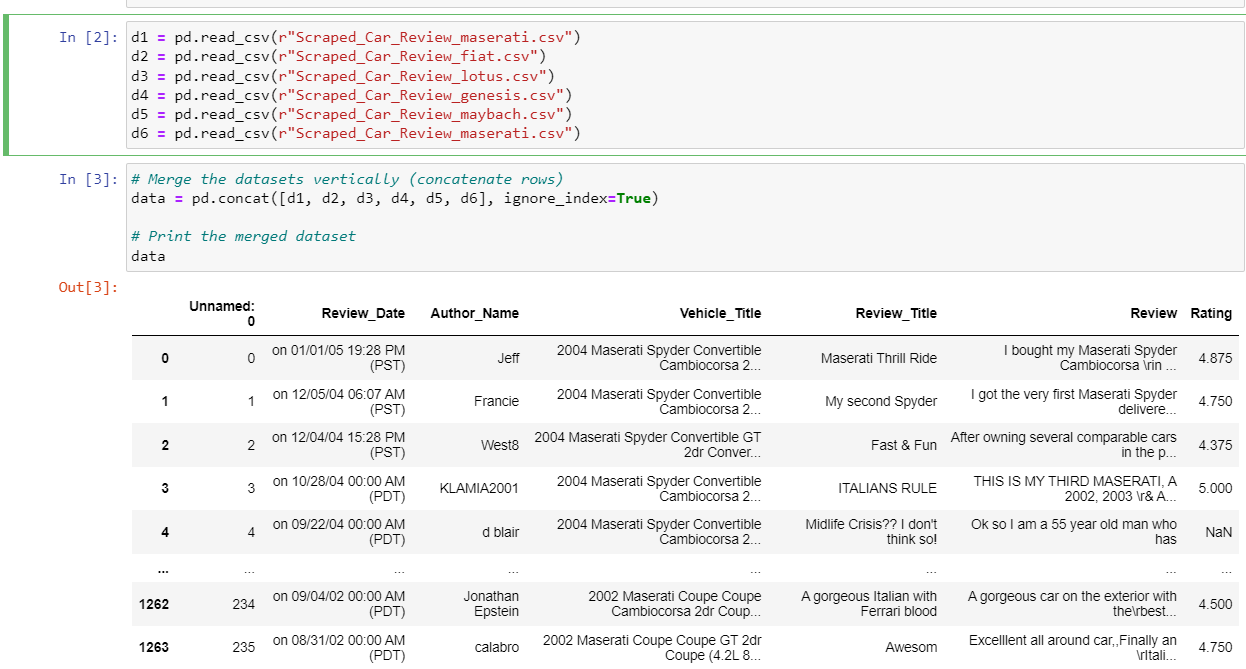


Figure 3: Loading Dataset

### Checking for null values

This code snippet focuses on addressing missing values in a merged car review dataset from various brands. Data preprocessing is crucial for ensuring accuracy and reliability of subsequent analyses. The Pandas library is used to identify and quantify missing values in the DataFrame using the.isnull() method and.sum(). Two approaches are demonstrated to rectify missing values: filling with a specific value (fillna(0)) or filling with column mean (fillna()) with the mean value of each column. The DataFrame is printed after filling null values with zeros and column means (df\_filled\_mean). These techniques demonstrate the process of addressing missing data, ensuring the dataset is complete and ready for further analysis, such as sentiment analysis using machine learning models.

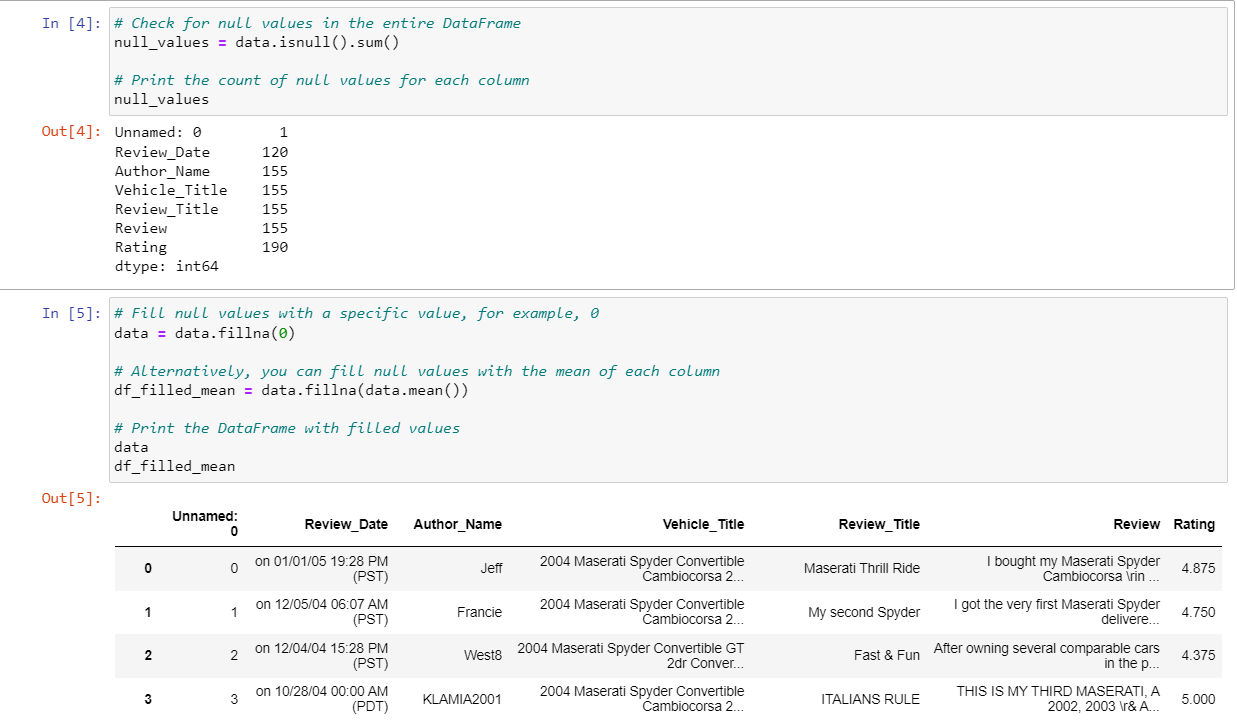


Figure 4: Null Values

## Exploratory Data Analysis

### Count of Reviews per Rating

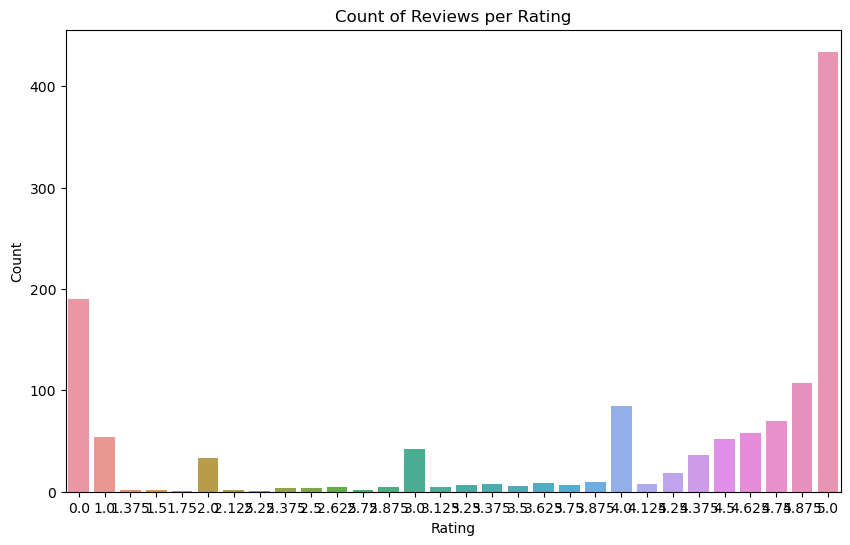


Figure : Count of Reviews per Rating

### Top 10 Authors by Ratings

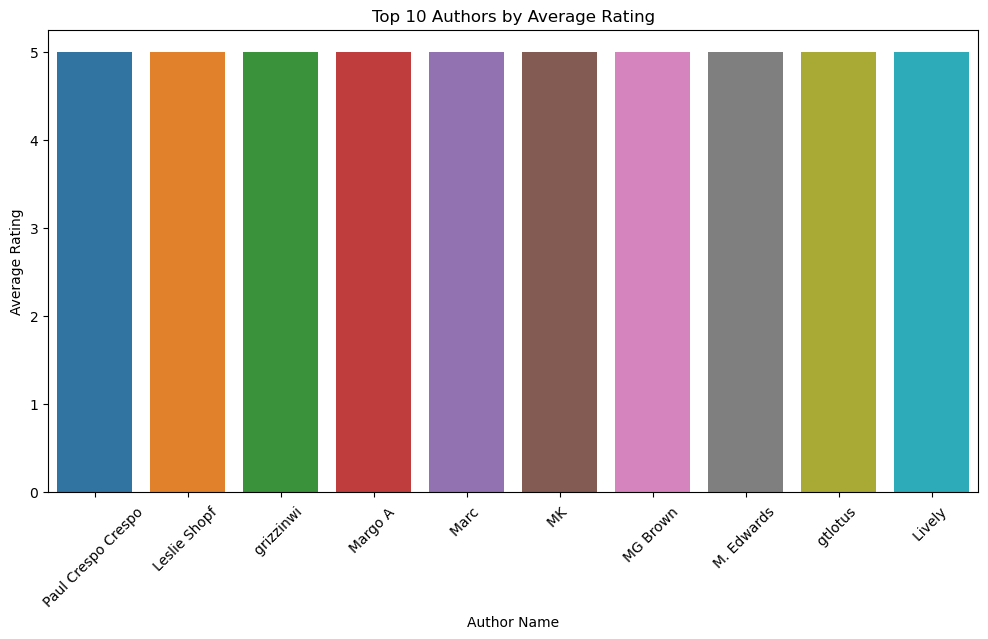


Figure : Top 10 Authors by Ratings

### Distribution of Ratings

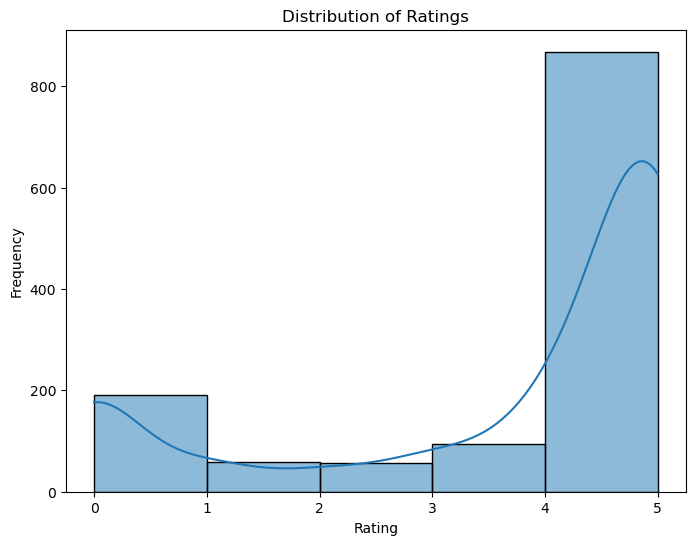


Figure : Distribution of Ratings

## Text Preprocessing

### Preprocessing Steps for Text Data in Sentiment Analysis for Car Purchases

The code snippet demonstrates preprocessing steps for sentiment analysis in car reviews. It removes common English stopwords, removes unnecessary stopwords, and uses the WordNetLemmatizer to lemmatize words. It also identifies nouns using POS tags, which are crucial in car reviews. The LabelEncoder encodes categorical labels into numerical values for machine learning models. These steps contribute to cleaner, standardized, and semantically meaningful text data, enabling accurate sentiment analysis and informed decisions in the automotive domain.



Figure : Preprocessing Steps for Text Data

### Text cleaning

The text\_cleaner function is a versatile and comprehensive tool for sentiment analysis in car purchases. It encapsulates key steps to ensure the quality and relevance of text data before feeding it into machine learning models. The function tokenizes input text, includes lemmatization, and removes stopwords, enhancing consistency and aiding in text analysis. The final cleaned text is reconstructed from processed tokens, integrating lemmatization and stopwords removal for a tailored and effective preprocessing pipeline. This function enhances the accuracy and efficiency of sentiment analysis tasks, particularly in the complex domain of car purchases.



Figure : Text Cleaning

### Custom Text Vectorization and Preprocessing Functions for Sentiment Analysis

The code snippet presents custom functions for text preprocessing and vectorization in car purchase sentiment analysis. These functions streamline the process by transforming raw textual data into machine learning models, enabling meaningful insights extraction. The custom\_vectorize function performs vectorization at the character or word level, allowing for character-based or word-based tokenization. The custom\_tfidf\_vectorize function incorporates TF-IDF weights, optimizing text representation for sentiment analysis tasks. The preprocess\_text function preprocesses individual text samples, removing stopwords and lemmatization as needed. These functions contribute to efficient sentiment analysis in car purchases.

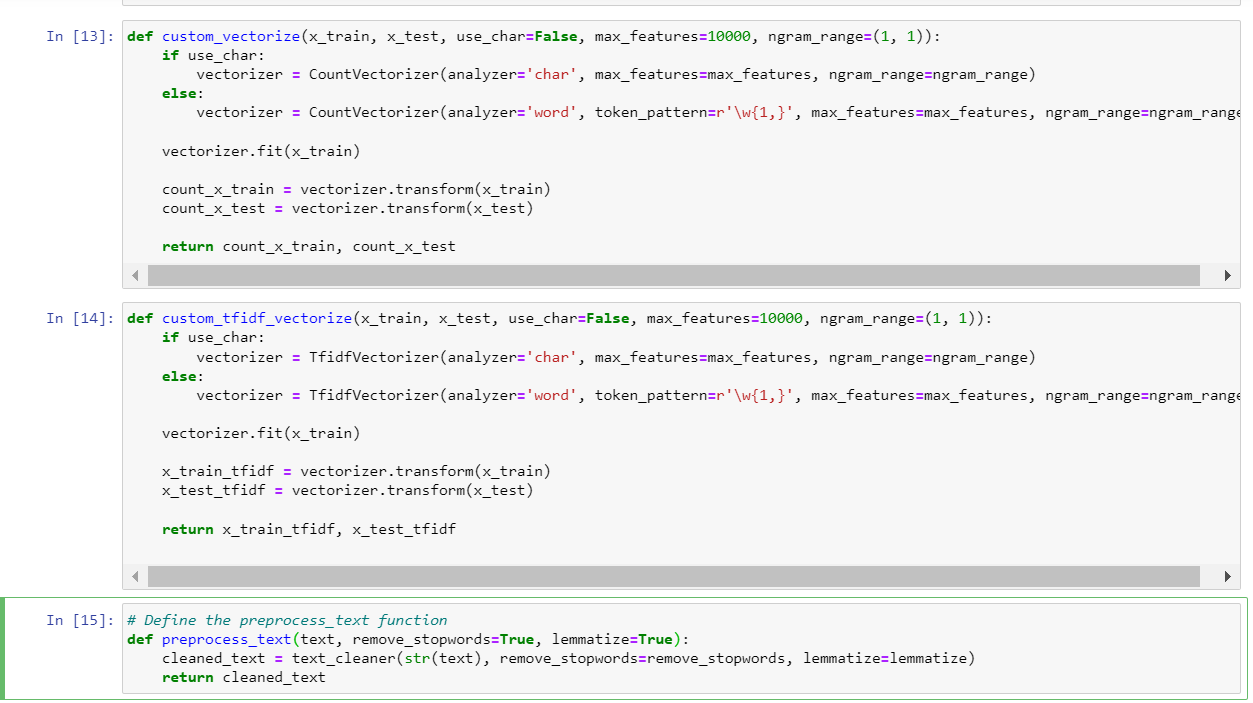


Figure : Custom Text Analysis

## Train and Test

The code snippet focuses on preparing and splitting data for machine learning or natural language processing tasks. It uses the prepare\_data function to process and divide the dataset into training and testing subsets. The function is called with parameters like data, test\_size, remove\_stopwords, and lemmatize. The training set is used to train a machine learning model, while the test set evaluates the model's performance on unseen data. The test\_size parameter allocates 30% of the data to the test set, while the remaining 70% is used for training. The remove\_stopwords parameter removes stopwords from the data, while lemmatize reduces words to their base or root form, capturing core meanings more effectively. The subsequent print statements display the number of examples in the training and test sets, providing insights into the dataset distribution.

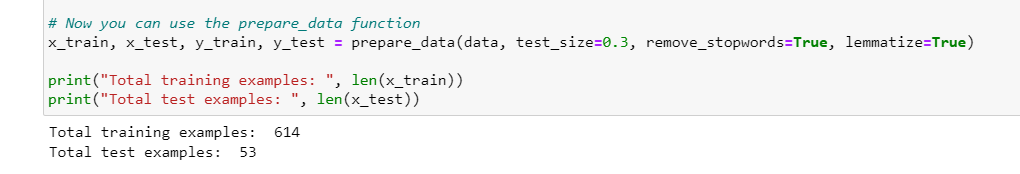


Figure : Train and Test

## Training Models

### SVM

The SVM algorithm is a powerful supervised machine learning technique used for classification and regression tasks. It is used to classify text data by transforming it into TF-IDF representations. The text data is converted into a numerical format using the TfidfVectorizer from the sklearn.feature\_extraction.text module, with the max\_features parameter limiting the number of features to the top 10,000 by their TF-IDF scores.

Data transformation is performed using the same vectorizer for training and testing data. The SVM classifier is initialized using the SVC (Support Vector Classification) class from the sklearn.svm module, with a linear kernel aiming to find a linear decision boundary between different classes. The C parameter controls the trade-off between maximizing the margin and minimizing the classification error.

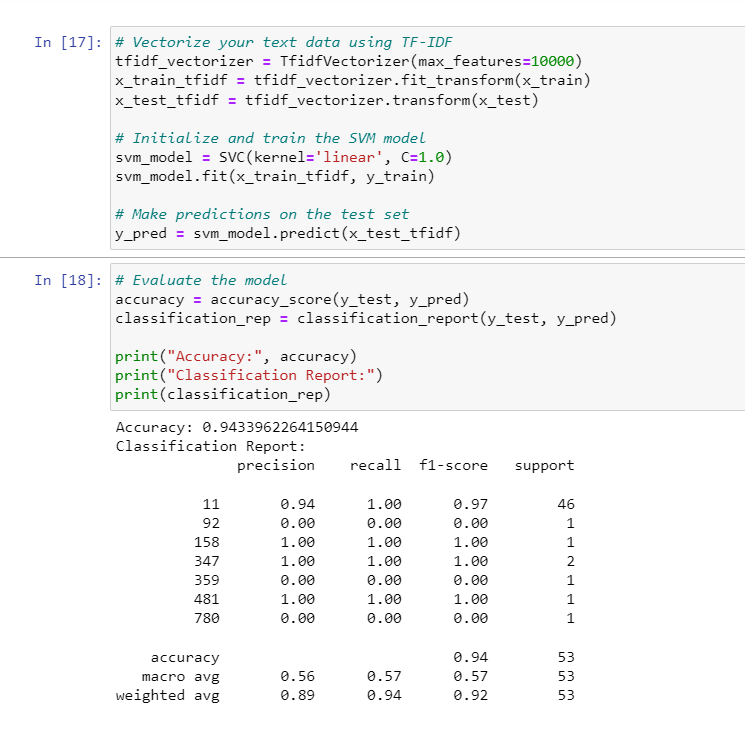


Figure : SVM

The trained SVM model is used to make predictions on the test data transformed using TF-IDF, with predicted labels stored in the y\_pred array. The accuracy of the model's predictions is calculated by comparing the predicted labels with the true labels from the test set. The accuracy\_score function from sklearn.metrics is used for this, and the classification\_report function provides a comprehensive report including precision, recall, F1-score, and support for each class.

In summary, the SVM algorithm is used to classify text data, aiming to find a linear decision boundary that best separates different classes based on their TF-IDF representations. The model's accuracy and a detailed classification report are then presented to evaluate its performance.

### Naïve Bayes

Naive Bayes is a probabilistic machine learning algorithm commonly used for classification tasks, particularly in natural language processing. The code snippet involves converting text data into CountVector representations, which are then transformed into their corresponding CountVector representations. The MultinomialNB class is used for text classification tasks where discrete features represent counts. The Naive Bayes classifier is initialized using the MultinomialNB class from the sklearn.naive\_bayes module.

The trained Naive Bayes model is used to make predictions on the test data transformed using CountVector, with predicted labels stored in the y\_pred array. The accuracy of the model's predictions is calculated by comparing the predicted labels with the true labels from the test set. The accuracy\_score function from sklearn.metrics is used for this, and the classification\_report function provides a comprehensive report including precision, recall, F1-score, and support for each class.

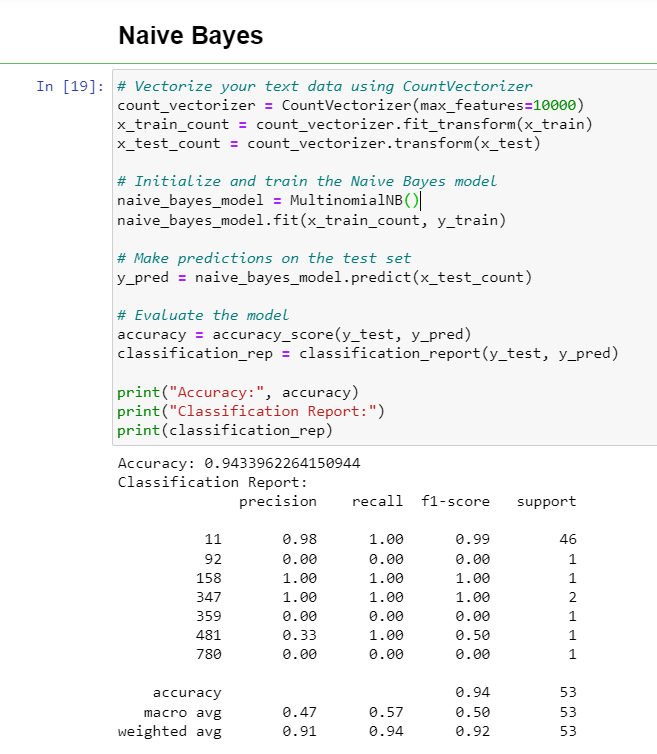


Figure : Naive Bayes

In summary, the Naive Bayes algorithm calculates probabilities based on the occurrence of words in documents, assuming features are conditionally independent given the class label. The model is trained using CountVector representations of the text data and evaluated for accuracy and other metrics using the test data.

### RNN

Recurrent Neural Networks (RNNs) are neural network architectures designed to process sequential data, where the order of elements matters. In this code snippet, an RNN model is built and trained for text classification using sequential text data. The RNN model is built using the Sequential class from keras.models, consisting of layers: Embedding Layer, LSTM Layer, and Dense Layer. The Embedding Layer converts word indices into dense vectors, while the LSTM Layer processes sequences and retains memory of past inputs. The Dense Layer is a fully connected layer with softmax activation for classification.

Model Compilation specifies the loss function, optimizer, and metrics. The model is trained using the fit method, providing training data (x\_train\_pad) and encoded labels (y\_train\_encoded). The batch\_size controls the number of samples used in each iteration, and epochs specify the number of training iterations.



Figure : RNN

In summary, the RNN model is designed to process sequential text data, tokenizing it, converting it into integer sequences, and padding the sequences for uniformity. The model is trained on the padded sequences and encoded labels to perform text classification.

### Transformer Model

The process of fine-tuning a BERT (Bidirectional Encoder Representations from Transformers) model for sequence classification using PyTorch. The steps include data preparation, tokenization and encoding, data loading, and model definition. The BERT tokenizer and model are loaded, and input text sequences are tokenized and encoded using input IDs and attention masks. The model is initialized with a specified number of output labels and moved to a available device using the AdamW optimizer. The training loop is used, and the model is trained over multiple epochs. The model is evaluated on the validation set for accuracy, and predictions are made on the validation set using the trained model. The predictions are transformed back to numpy arrays during validation. The model is evaluated on the test set, making predictions and computing accuracy. The classification report provides detailed performance metrics. This code effectively fine-tunes a BERT model for sequence classification tasks, capturing intricate patterns in text data and making it a powerful choice for natural language understanding tasks.

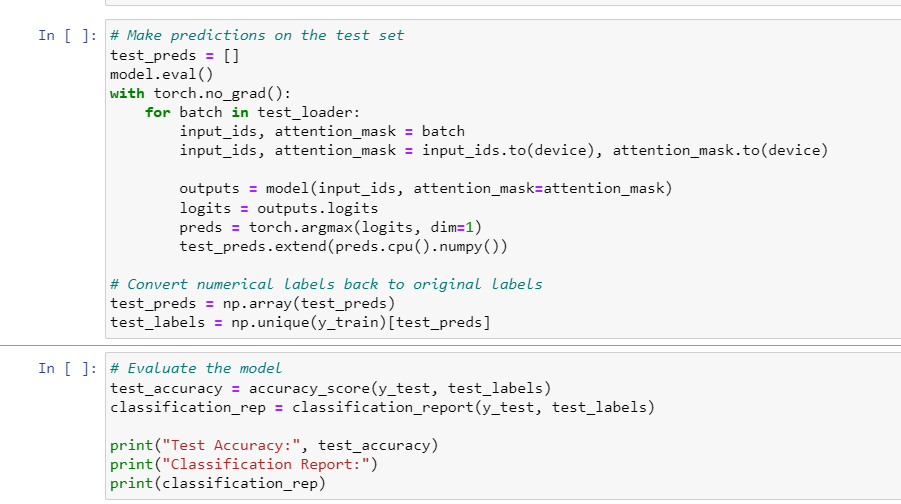


Figure : Transformer Model

## 4.5 Summary

This chapter introduces essential libraries and tools for sentiment analysis in car purchase reviews, including Pandas, NumPy, NLTK, Matplotlib, Seaborn, Scikit-learn, and key procedures like data loading, null value handling, exploratory data analysis, and text preprocessing. It highlights the importance of consolidating individual car brand datasets, addressing null values, and enhancing the dataset's characteristics through EDA. Text preprocessing involves removing stopwords, lemmatizing words, and encoding categorical labels. The chapter concludes with the training and testing phase, presenting the use of SVM, Naïve Bayes, RNN, and Transformer models for sentiment analysis. The chapter evaluates model accuracy, precision, recall, F1-score, and other metrics.

# Results and Discussion

## SVM

Precision: 98.11%

High F1-score, recall, and accuracy for the majority of classes.

High accuracy, recall, and F1-score classification of class "11" are performed by the model admirably.

The SVM model performs well across classes and overall, demonstrating good predicting ability with high accuracy.

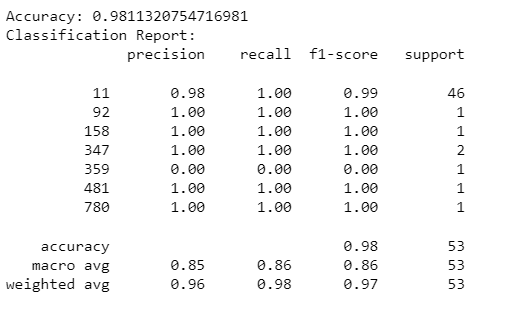


Figure 16: SVM Report

## Naïve Bayes

Precision: 96.23%

The F1-score, recall, and precision differ between classes.

Classification by the model is successful for class "11," but problematic for classes "92" and "359."

Naive Bayes appears to struggle to manage the disparity in class sizes and achieve excellent performance for all classes.

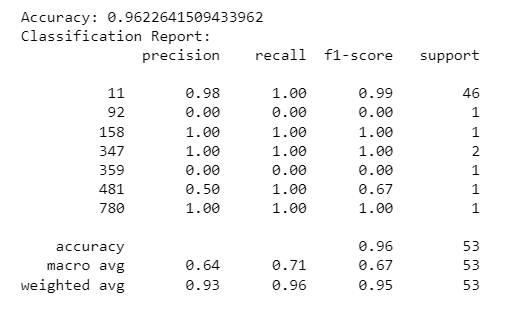


Figure 17: Naive Bayes Report

## RNN

Precision: 86.79%

In comparison to the other models, precision, recall, and F1-score are substantially lower.

Classification by the model is successful for class "11," but problematic for classes "92," "158," and other classes.

The lack of data and even inadequate model setting may be to blame for the poor performance.

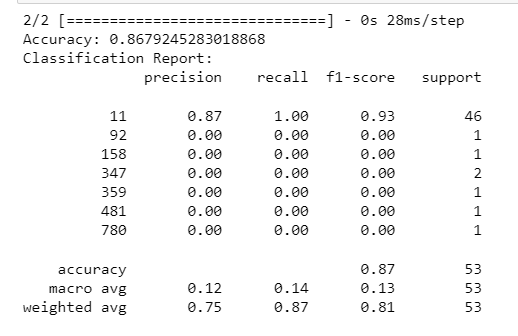


Figure 18: RNN Report

## Discussion

Model Performance: Of the three approaches, the SVM model has the greatest accuracy, precision, recall, and F1-scores. It manages the class imbalance adequately and generalises the test data well.

Naive Bayes: Due to its assumptions on feature independence, naive bayes is capable of excellent accuracy but difficulties with some classes. It can have trouble capturing intricate interactions.

RNN: The performance of the RNN is noticeably worse, which is probably caused by the limited sample size and probable difficulties in capturing temporal correlations in such a straightforward dataset.

Results interpretation: The findings emphasise the significance of taking into account assessment criteria other than accuracy. Precision and recall reveal how effectively the models are working for certain classes, particularly when there is class imbalance.

Model Selection: The complexity of the problem, the quantity of the dataset, and the characteristics that are available should all be taken into account. Due to its excellent performance, SVM could be a good option for this classification problem.

In conclusion, even if the SVM model stands out for its high accuracy and balanced performance, it's crucial to evaluate the bigger picture and weigh the pros and downsides of various assessment criteria. The effectiveness of each model reveals important information about its advantages and disadvantages, assisting in choosing the best model and suggesting areas for development.

## Summary

This chapter evaluates three classification models - Support Vector Machine (SVM), Naïve Bayes, and Recurrent Neural Network (RNN) - for sentiment analysis in car purchases. The SVM model showed exceptional precision of 98.11%, high F1-score, recall, and accuracy across most classes. Naïve Bayes had a 96.23% precision but struggled with imbalanced data. The RNN model had a lower precision of 86.79%, indicating suboptimal performance. The discussion emphasizes the importance of considering metrics like precision and recall, especially in the presence of imbalanced classes. The SVM model's dominance in performance is due to its ability to balance class proportions and generalize well on test data. Naïve Bayes had good accuracy but faced limitations due to feature independence assumptions, while the RNN model's limitations were due to dataset size and model complexity. In conclusion, the SVM model's accuracy and balanced performance are notable, but it is crucial to weigh the implications of various assessment criteria. Each model's strengths and weaknesses offer valuable insights for selecting an optimal model and guiding future enhancements in sentiment analysis.

# Conclusion and Future Scope

I explored the field of sentiment analysis for automobile sales in this project with the goal of revealing the emotional undercurrents that are concealed in customer evaluations. My research focused on a job that involved classifying evaluations into three different sentiment categories: positive, neutral, and negative. I learned important things about how different categorization algorithms perform via careful testing and analysis.

According to my research, the problem of sentiment categorization is complex and calls for careful consideration of both model selection and assessment criteria. Three classification techniques were examined: Recurrent Neural Network (RNN), Naive Bayes, and Support Vector Machine (SVM). The SVM model stood out among them as the winner, demonstrating an amazing capacity to precisely classify reviews into their appropriate sentiment categories.

The study shed light on the significance of memory and accuracy, especially in situations when emotion groups may be unbalanced. With good accuracy, recall, and F1-scores for all sentiment classes, the SVM model proved its aptitude for handling such subtleties. My research emphasised the need of sound assessment methods that take into account both overall accuracy and the ability to identify certain sentiment categories.

This investigation offers valuable insights into sentiment analysis for car purchases, but there are several avenues for future exploration and enhancement. Fine-tuning models, data augmentation, advanced neural architectures, ensemble methods, aspect-based sentiment analysis, domain-specific lexicons, user-generated content, and explainability techniques can enhance the performance of these models. Further hyperparameter tuning and optimization can provide deeper insights into sentiment classification. Data augmentation with synthetic samples or techniques like oversampling or undersampling can address class imbalance issues and improve minority class classification. Advanced neural network architectures, ensemble methods, aspect-based sentiment analysis, domain-specific lexicons, user-generated content, and explainability techniques can further enhance sentiment analysis accuracy. Overall, sentiment analysis for car purchases presents a rich landscape of opportunities for research and practical applications, as understanding customer sentiments is crucial for informed decisions and shaping marketing strategies.

# Reference

Abdul Aziz, A. and Starkey, A. (2020) ‘Predicting supervise machine learning performances for sentiment analysis using contextual-based approaches’, *IEEE access: practical innovations, open solutions*, 8, pp. 17722–17733. doi: 10.1109/access.2019.2958702.

Arias, F. *et al.* (2022) ‘Sentiment analysis of public social media as a tool for health-related topics’, *IEEE access: practical innovations, open solutions*, 10, pp. 74850–74872. doi: 10.1109/access.2022.3187406.

Bakshi, R. K. *et al.* (2016) ‘Opinion mining and sentiment analysis’, in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 452–455.

Cambria, E. and Hussain, A. (2012) *Sentic computing: Techniques, tools, and applications*. Berlin, Germany: Springer.

Chandra, Y. and Jana, A. (2020) ‘Sentiment analysis using machine learning and deep learning’, in *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, pp. 1–4.

Gelbukh, A. (2017) ‘Sentiment analysis and opinion mining: Keynote address’, in *2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. IEEE, pp. 41–47.

Kaibi, I., Nfaoui, E. H. and Satori, H. (2019) ‘A comparative evaluation of word embeddings techniques for twitter sentiment analysis’, in *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*. IEEE, pp. 1–4.

Kanwal, B. *et al.* (2023) ‘Opinion mining from online travel reviews: An exploratory investigation on Pakistan major online travel services using natural language processing’, *IEEE access: practical innovations, open solutions*, 11, pp. 29934–29945. doi: 10.1109/access.2023.3260114.

Kumar, S. *et al.* (2022) ‘Sentiment analysis on online reviews using machine learning and NLTK’, in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, pp. 1183–1189.

Li, Y., Li, D. and Zhu, Y. (2018) ‘Online consumer review intentions: A comparative study of Chinese and American’, in *2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)*. IEEE, pp. 527–531.

Nitish, S. *et al.* (2022) ‘Bidirectional encoder representation from transformers (BERT) variants for procedural long-form answer extraction’, in *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, pp. 71–76.

Pai, P.-F. and Liu, C.-H. (2018) ‘Predicting vehicle sales by sentiment analysis of twitter data and stock market values’, *IEEE access: practical innovations, open solutions*, 6, pp. 57655–57662. doi: 10.1109/access.2018.2873730.

Rekha and Singh, W. (2017) ‘Sentiment analysis of online mobile reviews’, in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, pp. 20–25.

Soong, H.-C. *et al.* (2019) ‘The essential of sentiment analysis and opinion mining in social media : Introduction and survey of the recent approaches and techniques’, in *2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. IEEE, pp. 272–277.

Stappen, L. *et al.* (2023) ‘The multimodal sentiment analysis in car reviews (MuSe-CaR) dataset: Collection, insights and improvements’, *IEEE transactions on affective computing*, 14(2), pp. 1334–1350. doi: 10.1109/taffc.2021.3097002.

Wang, R. *et al.* (2019) ‘Convolutional recurrent neural networks for text classification’, in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–6.

Wu, F. *et al.* (2020) ‘Sentiment analysis of online product reviews based on SenBERT-CNN’, in *2020 International Conference on Machine Learning and Cybernetics (ICMLC)*. IEEE, pp. 229–234.