National College of Ireland

Determining the Link Between Consumer Sentiments and Automobile Sales Through Sentiment Analysis

MSc Research Project
Data Analytics

Kartik Sharma
Student ID: 21125813

School of Computing
National College of Ireland

Supervisor: Prof.Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Kartik Sharma
Student ID:	21125813
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Prof.Christian Horn
Submission Due Date:	15/12/2022
Project Title:	Determining the Link Between Consumer Sentiments and Automobile Sales Through Sentiment Analysis
Word Count:	4698
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Kartik Sharma
Date:	30th January 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Determining the Link Between Consumer Sentiments and Automobile Sales Through Sentiment Analysis

Kartik Sharma
21125813

Abstract

Technology progresses every day, and millions of textual data points are created every second on the internet. Since the world is so interconnected today, people can use various web forums to find reviews of products before purchasing or renting them. Reviews and comments have a significant effect on a person's purchase decision. Consequently, it is essential for a company to understand its customers' behavior and attitudes toward its products and services and to use that information in sales, production, and marketing. This study examines how sentiments affect sales in the vehicle industry using the sentiment analysis component of Natural Language Processing (NLP). In the research the time series analysis to compare sales with sentiments is done to understand how both are connected, apart from this the lexical and pragmatic analysis of the reviews from the owners is done to understand their emotions about the vehicles they own. The Logistic Regression, Vader Sentiment Scoring, Roberta and neural network based models were created to predict the sentiments from a text review and the best model was selected based on the various evaluation metrics. Also it was found that the organizations shouldn't depend only on the ratings to decide whether customer actually liked or disliked the product or the services.

Keywords— Lexical Analysis, Pragmatic Analysis, Vader Sentiment Scoring, Roberta, Time Series Analysis, Natural Language Processing

1 Introduction

Governments and corporations have both employed consumer behavior and sentiment research when making key choices and focusing on high-yielding clients.Manually doing these studies was a time-consuming and difficult procedure, but with advances in NLP and neural network processing, as well as the usage of libraries such as Natural Language Toolkit (NLTK), these analyses can now be done with ease and high accuracy.
The project analyzes customer sentiment using a publicly available dataset of automobile evaluations. The findings of this investigation contribute to a better understanding of the link between customer mood and sales in the automobile industry.The findings will also assist the automotive sectors in making better judgments based on how their consumers perceive their products and services in order to attract new customers while maintaining existing ones.

1.1 Challenges with Sentiment Analysis

Sentiment Analysis is process of making machine understand the text and let machine decide which emotion is being conveyed, but the challenge arise when dealing with sarcasm, idioms, meaning of emotion icons(emojis) and tone of the text. Since human language is too ambiguous, it becomes tough for the machine to understand the patterns in it. Broadly the sentiment analysis can be broadly classified in to lexical analysis(understanding emotion from words as done by Vader sentiment scoring), syntactical analysis(understanding the relationship among words), semantic analysis(understanding meaningfulness of two words) and pragmatic analysis (letting machine understand the meaning of the sentence and the context, as done by RoBERTa pretrained model (refer to 3.3.3)).
For the research we used Pragmatic Analysis to achieve the objectives of the project.

1.2 Motivation

An organization's decision to market, produce, or sell a product is influenced by consumer sentiment. An automobile company may be able to use these feelings to determine what its clients feel about the amenities and services offered. In this way, the organization can assess which areas need further investment and improvement.

The case of General Motors Chevrolet India[1]

Chevrolet, a brand known for high-quality performance automobiles such as the Camaro, but the Indian market relied largely on a low-cost vehicle that can fit the entire family and is fuel-efficient. They introduced several automobiles into the market that were low on performance and quality but excellent on pricing, and the cars sold well. However, it proved impossible for General Motors to produce low-cost variants of automobiles that were selling successfully in other markets. Dealers also lost faith in Chevrolet India, which contributed to the company's demise. After the years of losses faced by Chevrolet India, general motors finally exited India in year 2017. Chevrolet first entered India in 2004, and it took them 13 years to realize that it was not the market they expected. They may have lowered their losses by leaving India earlier if they had done their Sentiment Analysis sooner.

1.3 Research Question and Objectives

How Artificial Neural Networks can be implemented for sentiment analysis in order to establish the link and predict consumer sentiments in relation to automobile sales ?

The goal of the project is to use the Artificial Neural Network aspect of machine learning to do sentiment analysis on consumer evaluations in order to determine how a consumer's general sentiment and emotion influences vehicle sales. Because of its parallel processing capabilities and higher accuracy than typical machine learning methods, the project employs the Artificial Neural Network methodology to fulfill its goals.

Objectives:

[1]https://www.motorbeam.com/general-motors-india-failure-5-reasons/

1. Do the ratings actually define the emotion in a review.
2. Do sales actually get affected by the general consumer sentiment.
3. Predicting the sentiment in a review using deep learning.

2 Related Work

2.1 Understanding Sentiment Analysis

The analysis of compound sentences is more complex than that of single sentences since real-time sentiment analysis will be conducted on reviews, which are usually compound sentences. To better comprehend compound sentence analysis, the project draws on the findings of this study (Savanur and Sumathi; 2018). In order to understand the emotions expressed in the phrases, the study uses conjunction analysis and feature-based analysis to place sentences in positive and negative categories.

The research (Shukri et al.; 2015) examines and compares the sentiments of customers between the various automobile brands. The study uses textual data collected from tweets and classify them based on the sentiment polarity. The research uses Naïve Bayes(NB) classifier for the achieving the goals, but the accuracy can be further increased by implementing RoBERTa process. The RoBERTa can help in generating more authentic results because it performs analysis based on the context and meaning of the sentence but the machine learning classifiers only predict sentiments based on lexical analysis.

Tough regression and classification based machine learning algorithms perform lexical analysis to predict the emotions is a text, but this is not accurate as the sentence can have sarcasm and tone variance, therefore lexical analysis suffers from these issues. Vader sentiment scoring(Veena et al.; 2021) uses similar approach, but to reduce the impact of issues with lexical analysis, it scores the text into 4 factors that are positivity, negativity, neutrality and compound. To have improved accuracy ,in the project we have used RoBERTa technique, because it does the pragmatic analysis rather than lexical analysis, hence takes the context of the sentence into consideration.

In order to better understand how customers behave, the project uses insights gained from this study (Sun et al.; 2019). This study delves into further detail on consumer behavior based on characteristics such as empathy, expectations, knowing a genuine customer, and comparing vendors. The research uses a data portrait paradigm to visualize data in order to make data-driven decisions.

People have online presence on social media at every second. Current world events are posted on social media and people express their views on the same. These data online can be used for companies benefit for finding new insights and user experience value. In the study(Ahn and Spangler; 2014) , they have used IBM CORPA to get data from online forms to get data about string “Sedan A buy” and “Sedan B buy”. Sentiment analysis and topical keyword analysis has been done to find the appropriate sentiments and words from the text scrapped. Before sales predictive model application an correlation testing has been performed to check auto correlation of columns. The model for sales prediction

used is ARIMA and sales prediction is mapped with the sentiments found by analysing the text. The Root mean squared error (RMSE) of Sedan B for sales based on history is 0.24 which is not up to the mark. This could have been overcome by having an original sedan A and sedan B sales data to compare it with.

2.2 Review of Sentiment Analysis in the Field of Sales

Vehicles have always been a luxury item, which has kept the automotive sector growing. Customers may now purchase motor cars ranging in price from low to high end value for personalized settings. Because there are various brands that offer automotive cars, the sales chart varies from brand to brand, even though the attributes of the automobile are the same. In the study (Shahid and Manarvi; 2009), they employed data mining and decision analysis to forecast sales patterns for various automobile cars. The car models have been differentiated on the basis of gasoline and compressed natural gas. As a result of the findings of this investigation, customers prefer dual fuel cars, as evidenced by the results.As a disadvantage, the sales trend should have been connected with the vehicle's feelings so that the cause for the decline or increase in sales could have been explained.

The majority of purchases in the internet era are directly or indirectly linked to online opinions about a product or brand. Analysis of online reviews and connecting them to sales becomes an essential part of the project. As a result, the project uses this study (K et al.; 2020) to gain a thorough understanding of internet sentiment analysis for projecting sales in the automobile sector. This study integrates a range of regression methodologies as well as sentiment analysis on data obtained from different vehicle forums on social media. This study aims to learn more about the link between sales and sentiments.

Nowadays, many companies compete for a client's business by offering the most competitive prices. As a result, it is critical for a company to understand how its consumers and future customers feel about the brand. This is in order to provide the most advantageous deal with the least or no financial losses to the firm and retain brand value. As a result, the project takes this part of the research (Athindran.N et al.; 2018) into account. This will enable it to gain a better understanding of customers' perceptions of different brands and their amenities. In this study, positive, neutral, or negative sentiment is classified using a Lexicon-based sentiment analysis approach and a Nave Bayes classifier.

Using sentiment analysis of Internet reviews and ratings, this research (Du and Yang; 2019) is based on sentiment analysis. Using ratings and reviews, this study seeks to develop a model that quantifies the emotional intensity of reviews. This will enable companies and organizations to evaluate consumers' opinions about their products and services. These insights are utilized in the project to better understand the impact of customer sentiment on an organization's marketing and sales strategy. In addition, they are used to tackle the problem of the emotional gap between ratings and reviews.

The Customer Relationship Management (CRM) approach is the foundation of this study (Zetu et al.; 2003) to gain insights into the elements that influence a person's decision to purchase a brand-new car. The method is based on survival analysis techniques and emphasizes the significance of marketing and customer behavior. The study demonstrates the significance of purchasing at the right time. The project might use the insights

gathered from this study to simulate customer behavior while considering scheduling and marketing elements. Numerous neural network and predictive analysis approaches may also aid in understanding customer behavior and feelings.

It was previously reported that (Zetu et al.; 2003) examined how CRM could help determine consumer behavior from a business perspective; this study, (Huang et al.; 2008), examines customer behavior by using decision tree algorithms based on age, education, income, gender, and car ownership. The purpose of this study is to determine how likely a person is to purchase a brand-new automobile and in what price range. The findings of this study may be utilized to choose the optimal algorithms for the project. In order to better understand how various car aspects influence purchase decisions, further sentiment analysis can be undertaken.

3 Methodology

The project implementation is based on the Knowledge Discovery in Databases (KDD) methodology. The Data considered is a large dataset consisting of reviews from consumers, therefore for analysis and model creation the KDD methodology is selected over Cross-Industry Standard Process for Data Mining(CRISP-DM) because unlike CRISP-DM, deep data processing and analysis can be done in the 9-Step KDD process. The KDD Process is preferred for vast datasets because deep data processing and analysis in an iterative way prove beneficial for Data Modelling to generate models with high accuracy.

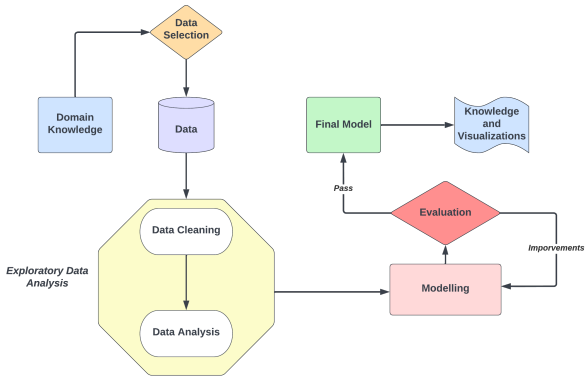


Figure 1: Project Methodology - Based on KDD process[2]

3.1 Domain Knowledge and Data

3.1.1 Domain Knowledge

Sentiment Analysis is one method in data analysis that combines human emotions with real-world quantitative data to analyze and infer how general human sentiment affects the political situation, business, and many other measurable elements in the world. For the study, car owner reviews are considered to determine how the combined consumer

sentiment affects automobile brand sales, and neural network models are trained to predict sentiment from the reviews.
The automobile industry's sales are heavily influenced by four factors: government policies, fuel prices, utility, and brand image. In today's internet age, with easy access to social media platforms, the client/consumer can read reviews from previous owners to make decisions, and any unfavourable variance in the above four characteristics might affect the reviews, hurting the customer's decision and, as a result, affecting sales.

3.1.2 Data

The Data is originally created by *Edmunds*(USA-based car selling business), and for the study, the data is taken from the *Kaggle.com* where it was publicly published. The Data consists of reviews from car owners in the USA from 2002 to 2018 in separate *(*comma separated value(.csv)) files for each car brand.
Original Dataset link: https://tinyurl.com/bdfjau8y

3.2 Data Preprocessing and Data Analysis

The complete Data processing and Analysis are done using Python programming language with Jupyter Notebook.

3.2.1 Data Preprocessing

Since every brand has its own csv file, therefore to begin the research all the csv files were combined into a single csv file using python and pandas. The *Author_Name* and *Vehicle_Title* columns were dropped as the study only focuses on the effect of sentiments on whole brand sales.
The Happiness scale given in Table 1 is added based on the rating column in the data frame. The rating was rounded off into a new column *Happiness_Level_RoundRatingBased* to have the absolute scale of happiness levels.

Table 1: Happiness Levels - Rating Based

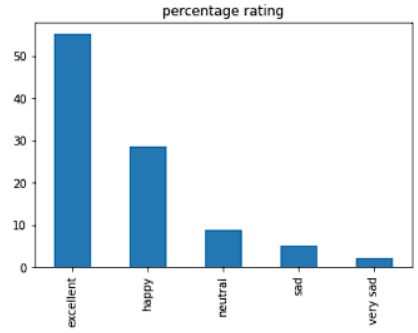
Rating	**Happiness Level**
1	Very Sad
2	Sad
3	Neutral
4	Happy
5	Very Happy

The *Review_Date* was processed to have the date in *dd//MM//yy* format. The text cleaning was performed on the *Reviews* column, where punctuation, special characters, and Emotion icons(emojis) were removed including converting the text to lowercase.

3.2.2 Data Analysis

For the initial analysis, the graph was plotted based on the happiness scale to get an idea of how the ratings are distributed in the dataset. Figure 2 shows that the data contains a large number of ratings over 3, therefore considering the ratings for the study can create

a bias in the analysis. It has also been seen that mistakenly a 5-star rating can have a bad review. To solve this issue we have focused the whole study on the sentiments derived from the consumer's reviews.




```
excellent    55.25
happy        28.62
neutral       8.78
sad           5.17
very sad      2.19
Name: Happiness_Level_RoundRatingBased, dtype: float64
```

Figure 2: Happiness Level Distribution in the Data

For finding out the relation between Sentiments and sales per brand, in the research line graphs based on sentiments from each brand are compared to each brand's sales data from year 2002 to 2018 and the patter are found which are discussed in section 5. To generate the emotions from sentence, the RoBERTa Pretrained model (discussed in section 4) is used and then average emotion per year per brand is used to plot the graphs. For comparison the sales data is sourced from *car sales base website*[3]

3.3 Modelling

Deep learning methods based on *Vader Sentiment Scoring* and *RoBERTa Pretrained model* as well as regression-based models, that is *Logistic Regression*, were utilized to create the models that can predict the emotion from the reviews in the project. Each modelling is discussed further in the section 4

3.3.1 Logistic Regression

Multinomial logistic regression, sometimes known as softmax regression because of the hypothesis function it makes use of, is a supervised learning method that may be used to solve a number of issues, including text classification. In this regression model, logistic regression is applied to classification problems with more than two possible outcomes(Ramadhan et al.; 2017).

3.3.2 Vader Sentiment Scoring

In Vader (Valence Aware Dictionary for Sentiment Reasoning), both the polarity (positive or negative) and the intensity (strong) of the sentiment are considered(Pai et al.; 2022). The Vader considers the words in the sentence and not the meaning of whole sentence. It scores the whole sentence into 4 parts that are negative, positive, neutral, and compound.

[3]https://carsalesbase.com/car-sales-us-home-main/

The positive score is the total average score of positive words in the sentence and similar calculation is for negative and neutral scores. The compound tells how much is the positivity of the sentence, higher the score more is the positivity.

3.3.3 RoBERTa Pretrained model

Similar to Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT pre-training Approach (RoBERTa) is a powerful pretrained language model(Liang et al.; 2022). With static masking, which BERT employs, the same portion of the text is hidden throughout each epoch. In contrast, RoBERTa employs dynamic masking[4], where different parts of the phrases are hidden during certain Epochs. The model gets strengthened as a result. The RoBERTa model are more reliable than the Vader scoring based models because it considers the meaning of whole sentence while scoring the sentence on 3 factors positive, negative and neutral score.

3.3.4 Neural Network based model

Neural networks[5] are artificial intelligence technologies that train computers to interpret data in a similar manner to how the human brain does. A deep learning machine learning approach that uses layered nodes, similar to a human brain, to learn. By employing adaptive mechanisms, computers are able to continually learn from their mistakes. In order to solve complex issues such as summarizing papers or recognizing faces, artificial neural networks strive to improve accuracy.
Using natural language to process human-created materials is known as natural language processing (NLP). In order to extract insights and meaning from text data and documents, computers use neural networks.

3.4 Model Evaluation

To evaluate[6] the models, in the project the following metrics are considered:

3.4.1 Accuracy Score

An accuracy measurement shows how often a classifier gets the right answer. In forecasting, accuracy can be defined as the ratio of correct predictions to total predictions. The accuracy for the models in the project are calculated by taking in the consideration the test data.
The formula for calculating the accuracy score is give in figure 3

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Figure 3: Formula to calculate the accuracy for the classification model[3.4]

[4]https://towardsdatascience.com/exploring-bert-variants-albert-roberta-electra-642dfe51bc23
[5]https://aws.amazon.com/what-is/neural-network/
[6]https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/

- TP - True Positive
- TN - True Negative
- FP - False Positive
- FN - False Negative

3.4.2 Sensitivity (Recall)

Recall measures how many positive cases our model correctly predicted. When False Negative exceeds False Positive, it is a useful indicator The formula for calculating the Sensitivity is give in figure 4

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

Figure 4: Formula to calculate the Sensitivity for the classification model[3.4]

3.4.3 Precision

Using precision, we can determine how many of the predicted positive outcomes actually took place. A high degree of precision is especially important when False Positives are more of a concern than False Negatives. The formula for calculating the Precision is give in figure 5

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Figure 5: Formula to calculate the Precision for the classification model[3.4]

3.4.4 ROC Curve

Receiver operator characteristic (ROC)[3.4] measures the difference between true positives and false positives at different threshold levels and separates 'signal' from 'noise.
It measures how well a classifier can differentiate between classes by measuring the Area Under the Curve (AUC).

3.4.5 Confusion Matrix

The confusion matrix measures the performance of machine learning classification tasks that have two or more output classes. This table contains a mixture of expected and actual data. The Calculation of the Confusion Matrix is given in figure 6

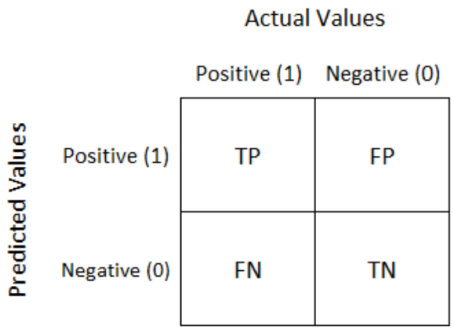


Figure 6: Confusion Matrix for the classification model[3.4]

3.5 Design Specification

The following is the architectural design of the project:

1. The data is downloaded from kaggle and stored in the local storage
2. The data is preprocessed, all the null values, special characters, stopwords, emotion icons were removed and the text data was changed to lower case. Apart from this the columns not required for the research were removed from the dataset.
3. The Exploratory Data Analysis is done on the data understand it better and to generate deeper insights and patterns.
4. To perform sentiment analysis and create models that can predict sentiments from the text, three aproaches are taken and best approach after evaluation was combined with Neural Network to generate final model. The 3 approaches are given below:
    - Logistic Regression
      The Machine learning algorithm for classification that use Lexical Analysis for sentiment prediction.
    - Vader Sentiment Scoring
      The Deep Learning algorithm that use Lexical Analysis for sentiment prediction.
    - RoBERTa pretrained model
      Model trained on twitter textual tweets that uses Pragmatic Analysis for the sentiment prediction.

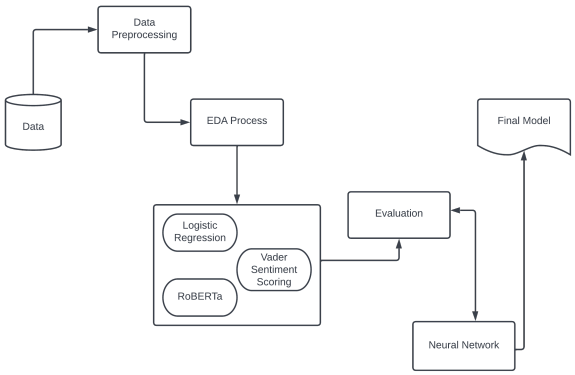


Figure 7: Project Design

4 Implementation

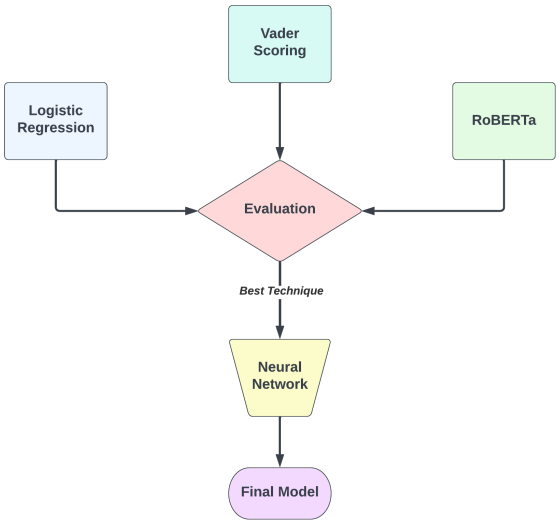


Figure 8: Project Implementation to generate sentiment predicting model(Self Created)

4.1 Logistic Regression

To apply Logistic Regression for the multi-class classification of reviews based on the emotions in the project, the data was split with 80% for the training and 20% for the testing dataset.
The *Term Frequency Inverse Document Frequency*[7] *Vectorizer*(TfidfVectorizer) is used to convert the text into machine understandable combination of numbers which in turn can be used to fit the machine learning algorithms.
The SAGA solver is used for Logistic Regression as it is fast for large datasets where multi-class problems are needed to be solved. The dataset for the research is both large and has multiple classes of emotions.

[7]https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a

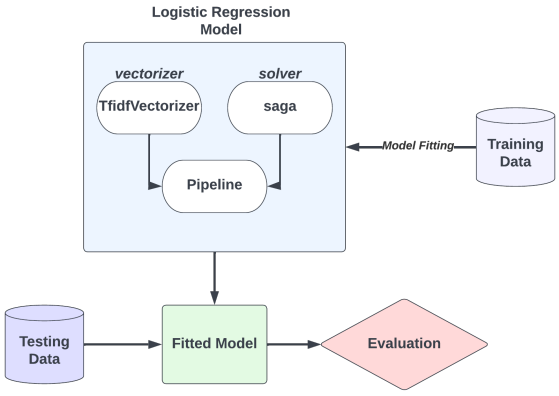


Figure 9: Implementation using Logistic Regression(Self Created)

4.2 Vader Sentiment Scoring

The next technique used in the research is the Vader Sentiment Scoring, it takes each word in the review, and calculates the average of the negative(neg), positive(pos), neutral(neu) and a compound score of the review based on each word in the sentence and not considering the whole meaning of the sentence.
To calculate the score the Sentiment Intensity Analyzer(sia) from the Natural Language Toolkit (NLTK) library.

```
In [17]: from nltk.sentiment import SentimentIntensityAnalyzer
         from tqdm.notebook import tqdm

         sia = SentimentIntensityAnalyzer()

In [18]: sia.polarity_scores('I am somewhat happy ')
Out[18]: {'neg': 0.0, 'neu': 0.37, 'pos': 0.63, 'compound': 0.5279}
```

Figure 10: Using SIA for Vader sentiment scoring (Self Created)

4.3 RoBERTa Pretrained model

For the project, the RoBERTa pretrained model is preferred because it does the sentiment scoring more confidently than the Vader sentiment scoring, refer to figure 11. Apart from this RoBERTa is much better than Vader because it considers the meaning and context of the sentence rather than only words.

- neg - average negative score of sentence using Vader Sentiment Scoring
- neu - average neutral score of sentence using Vader Sentiment Scoring
- pos - average positive score of sentence using Vader Sentiment Scoring
- Rob_neg - average negative score of sentence using RoBERTa PreTrained Model
- Rob_neu - average neutral score of sentence using RoBERTa PreTrained Model
- Rob_pos - average positive score of sentence using RoBERTa PreTrained Model

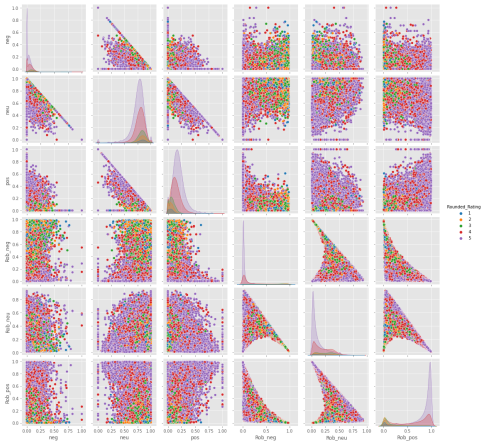


Figure 11: Roberta Scoring Vs Vader Scoring(Self Created)

RoBERTa model is pretrained on the *Twitter* tweets for sentiment analysis. The RoBERTa model was used to extract sentiment scores from each sentence, based on the average scores the sentiments were classified into positive, negative and neutral in a new column in the dataset, as shown in the figure 12.
This technique is known as Transfer learning[8], in which knowledge gained from solving one problem is used to solve the other similar problem.

	id	neg	neu	pos	compound	Rob_neg	Rob_neu	Rob_pos	index	Review_Date	Review_Title	Review	Rating	Brand
0	0	0.000	0.825	0.175	0.7430	0.003627	0.029560	0.966813	0	2005-10-13	Great delivery vehicle	It's been a great delivery vehicle for my caf...	4.625	dodge
1	1	0.125	0.776	0.099	-0.4793	0.858091	0.123257	0.018652	1	2005-07-17	Disappointmnet	Bought this car as a commuter vehicle for a v...	2.125	dodge
2	2	0.000	0.707	0.293	0.8519	0.002176	0.013060	0.984764	2	2002-07-16	Sweet van	This van rocks its the best, lots of vroom....	5.000	dodge
3	3	0.000	0.570	0.430	0.8689	0.002926	0.023611	0.973463	3	2007-12-29	Keven Smith	Great work vehicle. Drives nice, has lots of ...	4.500	dodge
4	4	0.126	0.757	0.115	-0.4201	0.500820	0.350322	0.148858	4	2005-02-09	Not what Dodge used to be	Good solid frame and suspension. Well equipp...	2.875	dodge

Figure 12: Data with Roberta Scoring(Self Created)

4.4 Neural Network based model

Before proceeding for the neural network, the new data frame, refer to figure 13 was created with *cleaned_Reviews* as independent variable X and the *Rob_Emotion* as the target variable Y. The *Rob_Emotion* is the column which was created based on the average scores from the RoBERTa technique for each sentence. This column tells if the sentence is overall positive, negative or neutral. The column with dummy variables where 0 for negative, 1 for neutral and 2 for positive is created for the machine learning. The stopwords (for, if, have, it, and many others) were identified in the reviews usng *stopwords* from nltk corpus, and were removed from the reviews using *PorterStemmer* from nltk.stem.porter, and the process is known as stemming[9]. The *CountVectorizer*[10] from sklearn.feature_extraction.text is then used to transform the reviews it into the machine

[8]https://en.wikipedia.org/wiki/Transfer_learning
[9]Stemming is the process to extract the most basic form of the word by removing the suffix and affix.

readable format. The sequential[11] neural network model having 3 dense[12] layers with 12, 8, and 8 output size respectively with relu[13] , relu and softmax[14] as the activation functions respectively is compiled with the loss as *sparse_categorical_crossentropy* and optimizer as *adam*[15] is created. The data was split, where the training data was the 70% of the whole data. The model was fitted on the training data with 10 epochs and 10 as the batch size.

	cleaned_Reviews	Rob_Emotion
0	its been a great delivery vehicle for my cafe...	positive
1	bought this car as a commuter vehicle for a v...	negative
2	this van rocks its the best lots of \rroom i ...	positive
3	great work vehicle drives nice has lots of ro...	positive
4	good solid frame and suspension well equippe...	negative
...	...	...
212847	i have owned the tiguan for a year and month...	positive
212848	now had months with suv nice suv handles we...	negative
212849	smaller dimensions and my driving experience ...	positive
212850	we have had our tiguan for a month prior to ...	positive
212851	the tiguan s can be had for a reasonable and...	positive

212852 rows × 2 columns

Figure 13: Data Frame with reviews and respective sentiments(Self Created)

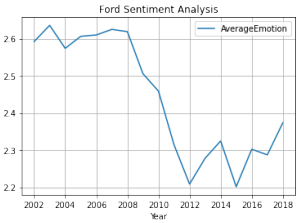
5 Results and Evaluation

5.1 Results from Data Analysis

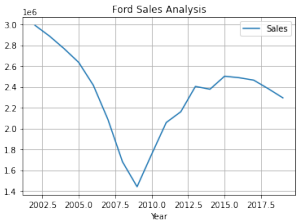
The sentiments are compared to the sales of respective brands and the results are explained below.

After analyzing the figures 14, 15, 16 and 17 a pattern was found that the sentiments were on the lower levels (less than 2.5), more near to negative and the auto mobile sales were also falling during the period between year 2005 to 2012. Doing further research, it was found that at same time the government policies regarding the emissions were changed and implemented rigorously, the oil prices were increasing and the customers were look for cheap fuel efficient vehicles, this period was also know as the *Automobile Industry Crisis*[16] and this was further affected by Great Recession that had the severe impact on U.S. economy.
The customers shifted to the cheaper alternatives provided by the Japanese makers like

[10]tokenizes the words in text as 0s and 1s, and then the text is represented as sparse matrix.
[11]it helps in creating the model layer by layer
[12]helps in defining each neuron in each layer of the network
[13]Rectified Linear Unit, has the advantage that it doesn't activates all neurons together, therefore less resource consumption and faster computation
[14]normalizes the network output across projected output categories, normalize each output class's output to a probability distribution
[15]one of the best optimizers when working with the sparse data, its adaptive learning rate is best for sparse data
[16]https://en.wikipedia.org/wiki/Automotive_industry_crisis_of_2008%E2%80%932010

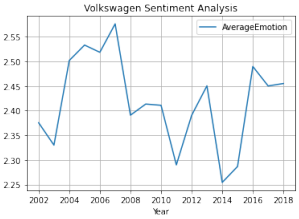


(a) Ford Customer Sentiments(Self Created)

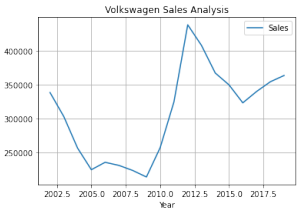


(b) Ford Sales(Self Created)

Figure 14: Ford Sentiment Vs Sales

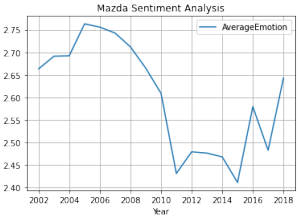


(a) Volkswagen Customer Sentiments(Self Created)

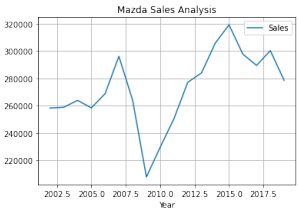


(b) Volkswagen Sales(Self Created)

Figure 15: Volkswagen Sentiment Vs Sales

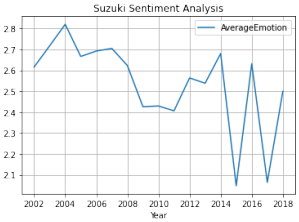


(a) Mazda Customer Sentiments(Self Created)

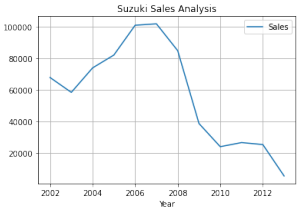


(b) Mazda Sales(Self Created)

Figure 16: Mazda Sentiment Vs Sales



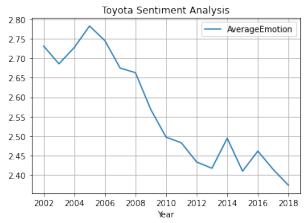
(a) Suzuki Customer Sentiments(Self Created)



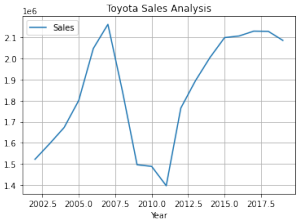
(b) Suzuki Sales(Self Created)

Figure 17: Suzuki Sentiment Vs Sales

Toyota. Though Toyota sales fell initially, but due to their marketing and making cheaper cars, they recovered exponentially. Clearly the sales increased but the customers were not happy with the quality and performance of the cars, which can be seen from the Toyota Customer Sentiments plot. Toyota must have decreased their cars quality and performance in order to provide cheap fuel efficient cars.



(a) Toyota Customer Sentiments(Self Created)



(b) Toyota Sales(Self Created)

Figure 18: Toyota Sentiment Vs Sales

5.2 Evaluation

5.2.1 Experimenting with Logistic Regression

For evaluating we used accuracy, precision and Recall scores from the *sklearn.mertics* ,refer to figure 19. The model had low accuracy of 65% with a low precision score, therefore the model was rejected. Clearly, from figure 20, the model was unable to predict the negative emotion in the sentence.

```
Model Evaluations for  Logistic Regression
Accuracy :  0.6511801126576975
Precision :  0.7072135037289162
Recall :  0.6511801126576975
              precision    recall  f1-score   support

   excellent       0.75      0.89      0.81     23700
       happy       0.51      0.41      0.45     12405
     neutral       0.37      0.31      0.34      3755
         sad       0.38      0.26      0.31      2190
    very sad       0.43      0.16      0.23       912

    accuracy                           0.65     42962
   macro avg       0.49      0.41      0.43     42962
weighted avg       0.62      0.65      0.63     42962
```

Figure 19: Classification Report Logistic Regression model(Self Created)

```
Test_Review = ['I am not happy to buy this car']
Test_prediction = model_lr.predict(Test_Review)

print(Test_prediction)
```

```
['excellent']
```

Figure 20: Sample Text test on Logistic Regression model(Self Created)

5.2.2 Experimenting with Vader Sentiment Scoring

The SIA performed well is extracting the sentiments from the reviews, as when compared to the ratings by the consumer the sentiments are sentiment scores are well aligned as can be seen in figure 21 and figure 22, for example, refer to figure 22 the positive graph (left graph), the positivity is increasing with the better ratings

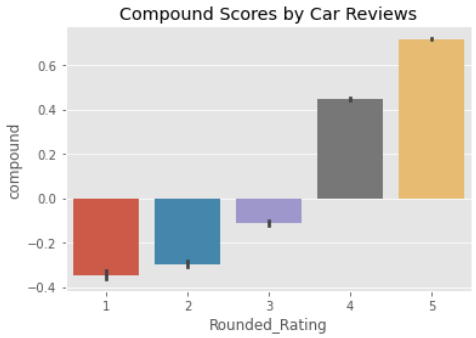


Figure 21: Compound Vs ratings (Self Created)

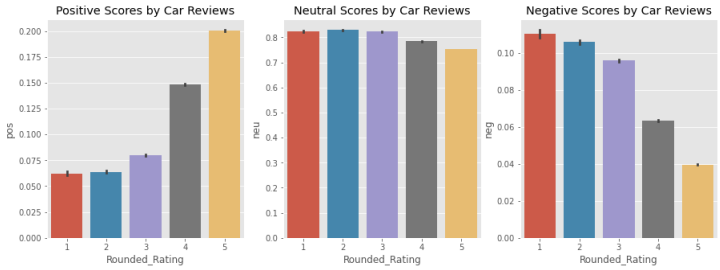


Figure 22: Positive neutral and negative scoring Vs ratings (Self Created)

5.2.3 Experimenting with RoBERTa Pretrained Model

From the figure 23, it can be proven that the ratings alone cannot be trusted, clearly the model classified the review as positive when the rating given was low. The results showed about 7% of the negatively rated reviews were actually positive and 55% of the high rated reviews are actually negative.

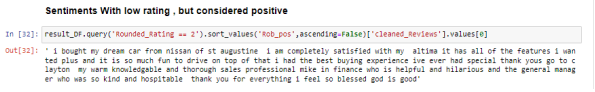


Figure 23: Data with Roberta Scoring(Self Created)

5.2.4 Experimenting with Neural Network based Model

The Neural network was fitted with different epochs and the constant batch size of 10, the following observations were made, refer to the table 24. Though with training accuracy

was less with the 10 epochs, and loss was highest, as compared to the other values of epochs, the test accuracy was higher. Other epochs were also having the similar accuracy, but considering the lesser run time and low utilization if resources, the model fitted with epochs = 10 is preffered. The figure 25 is the illustration how the the neural network

epochs	loss (%)	Train_Accuracy(%)	Test_Accuracy(%)
10	17	95	86
30	9	97	83
100	2	99	82

Figure 24: Model comparison with respect to epochs(Self Created)

based model was able predict the sentiment from the text take from the source unknown to the model.

```
In [84]: text = 'Sound quality is very bad. I returned the product'
         text = dataPreprocessing(text)
         array = cv.transform([text]).toarray()
         prediction = model.predict(array)
         a=np.argmax(prediction,axis=1)
         lable_encoder.inverse_transform(a)[0]

         1/1 [==============================] - 0s 38ms/step

Out[84]: 'negative'
```

Figure 25: Model predicting the sentiment from unknown text(Self Created)

6 Conclusion and Future Work

The research is focused on determining the effect of human sentiments on the sales, and developing a model that can predict the emotion from the text. While doing the project, it was found that the rating only, cannot be trusted because human mind can fail to determine their sentiments in numbers, therefore to better understand the sentiments, analysing the comments and reviews is a better approach. To perform the sentiment analysis we used Logistic Regression, Vader Sentiment Scoring and RoBERTa technique, and RoBERTa performed better in determining the sentiments, this is because logistic regression and Vader Sentiment Scoring both uses lexical analysis approach but RoBERTa perform pragmatic analysis to understand the meaning and context to the sentence. From the research, it proved that ups and downs in human sentiments whether due to financial conditions, government policies or mere utility can truly have an affect on a organization's sales chart.

In future, more data can used to train same model by increasing the computer resources to increase the accuracy of the sentiment prediction. The model testing can be done on various data sets in order to confirm the model's accuracy. The model was trained to classify the sentiments into positive, negative and neutral, but using the *ncrlex* library, the text can be classified based on more number of emotions like empathy, contempt,

guilt, gratitude and many others, thereofore using this the trained model will be able to perform much deeper analysis and thus the accuracy of the results can be improved.

7 Acknowledgement

My sincere thanks to the project supervisor Prof.Christian Horn, the whole faculty of Research In Computing (National College of Ireland) and to all my friends who supported me to understand the project and its documentation in the best ways possible.

References

Ahn, H.-I. and Spangler, W. S. (2014). Sales prediction with social media analysis, *2014 Annual SRII Global Conference*, pp. 213–222.

Athindran.N, S., Manikandaraj.S and Kamaleshwar.R (2018). *Comparative Analysis of Customer Sentiments on Competing Brands using Hybrid Model Approach*.

Du, Y. and Yang, L. (2019). A sentiment measurement model for online reviews of complex products, Institute of Electrical and Electronics Engineers Inc., pp. 199–202.

Huang, L., Zhou, C. G., Zhou, Y. Q. and Wang, Z. (2008). Research on data mining algorithms for automotive customers' behavior prediction problem, pp. 677–681.

K, P. S., Vikyhat, S., Pranav, S. and Abhishek, Y. (2020). *Sales Prediction using Online Sentiment with Regression Model*.

Liang, J., Martinez, A. and Morita, H. (2022). Language model adaptation for downstream tasks using text selection, *2022 4th International Conference on Natural Language Processing (ICNLP)*, pp. 320–323.

Pai, A. R., Prince, M. and Prasannakumar, C. V. (2022). Real-time twitter sentiment analytics and visualization using vader, *2022 2nd International Conference on Intelligent Technologies (CONIT)*, p. 3.

Ramadhan, W., Astri Novianty, S. and Casi Setianingsih, S. (2017). Sentiment analysis using multinomial logistic regression, *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)*, pp. 46–49.

Savanur, S. R. and Sumathi, R. (2018). Feature based sentiment analysis of compound sentences, *2017 2nd International Conference On Emerging Computation and Information Technologies, ICECIT 2017* .

Shahid, S. and Manarvi, I. (2009). A methodology of predicting automotive sales trends through data mining, *2009 International Conference on Computers Industrial Engineering*, pp. 1464–1469.

Shukri, S. E., Yaghi, R. I., Aljarah, I. and Alsawalqah, H. (2015). Twitter sentiment analysis: A case study in the automotive industry, *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pp. 3–4.

Sun, J., Zhao, H., Mu, S. and Li, Z. (2019). Purchasing behavior analysis based on customer's data portrait model, Vol. 1, IEEE Computer Society, pp. 352–357.

Veena, G., Vinayak, A. and Nair, A. J. (2021). Sentiment analysis using improved vader and dependency parsing, *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pp. 2–4.

Zetu, D., Cheng, J., Jay, J. M. and Lo, B. (2003). When will a consumer consider buying another car?, pp. 530–534.