REVA UNIVERSITY
Bengaluru, India

A Project Report on

Comparison Metrics of Buyer's Voice on Indian SUVs - Sentiment Analysis

Submitted in partial fulfilment for award of degree of

BA05

In **Business Analytics**

Submitted by

Anand Limbare
R19DM001

Under the Guidance of

Sandeep Giri

Founder at CloudxLab

REVA Academy for Corporate Excellence
REVA University
Rukmini Knowledge Park, Kattigenahalli,
Yelahanka, Bangalore – 560064

October, 2020

REVA
UNIVERSITY
Bengaluru, India

Candidate’s Declaration

I, **Anand Limbare** hereby declare that I have completed the project work towards the first year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled ‘**Comparison Metrics of Buyer's Voice on Indian SUVs - Sentiment Analysis**’ under the supervision of **Sandeep Giri, Founder at CloudxLab**. This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year **2020**.

Place: Bengaluru
Name of the Student: Anand Limbare

Date: 29/10/2020
Signature of Student:

REVA
UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the Project work entitled ‘**Comparison Metrics of Buyer's Voice on Indian SUVs - Sentiment Analysis**’ carried out by ‘**Anand Limbare**’ with ‘**R19DM001**’, is a bonafide student of REVA University, is submitting the first year project report in fulfilment for the award of **BA05** in Business Analytics during the academic year **2020**. The Project report has been tested for plagiarism, and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of PROJECT work prescribed for the said Degree.

Signature of the Guide	Signature of the Director
Sandeep Giri	Dr. Shinu Abhi
Founder at CloudxLab	Director (RACE)

External Viva

Names of the Examiners

1. <Name> <Designation> <Signature>
2. <Name> <Designation> <Signature>

Place: Bengaluru

Date: 29/10/2020

REVA
UNIVERSITY
Bengaluru, India

Acknowledgement

I would like to express my deepest thanks to my mentor Mr. Sandeep Giri for his immense knowledge that helped my project find a structure. I would also like to thank Dr. Shinu for reviewing my report and providing me with useful insights for improvement. I would also thank my friends at RACE who have assisted me in building my motivation that required me to complete this report.

And most of all thanks to my parents who supported me from the beginning of this program, believed in me and were the backbone of my development.

I am thankful for and would like to acknowledge Hon’ble Chancellor, Dr. P Shayma Raju, Vice Chancellor, Dr. K. Mallikharjuna Babu, and Registrar, Dr. M. Dhanamjaya, for giving me a chance to enrol in this program with which I have gained immense knowledge that cannot be quantified.

Place: Bengaluru

Date: 29/10/2020

REVA
UNIVERSITY
Bengaluru, India

Similarity Index Report

This is to certify that this project report titled ‘**Comparison Metrics of Buyer's Voice on Indian SUVs - Sentiment Analysis**’ was scanned for similarity detection. Process and outcome is given below.

Software Used: Turnitin

Date of Report Generation: 30/10/2020

Similarity Index in %: 9%

Total word count: 6904

Name of the Guide: Sandeep Giri

Place: Bengaluru Anand Limbare

Date: 29/10/2020

Verified by:

Signature

Dr. Shinu Abhi,

Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	StanfordNLP	Stanford Natural Language Processing
2	SUV	Sports Utility Vehicle
3	LDA	Latent Dirichlet Allocation
4	GDP	Gross Domestic Product
5	CRISP – DM	Cross-industry standard process for data mining
6	IE	Internet Explorer
7	URL	Uniform Resource Locator
8	CSV	Comma Separated Values
9	GST	Goods and Service Tax
10	IP	Internet Protocol
11	EDA	Exploratory Data Analysis
12	NLTK	Natural Language Toolkit
13	KM	Kilometre

List of Figures

List of Tables

Abstract

The better the reviews, the more we trust a business. This project aims to solve two most common industry problem in the Indian Auto Industry, choosing a well-crafted product for a consumer and making necessary modifications on the coming iterations for manufacturers on their components. Unadulterated reviews are often found in review forums where the aggregator's or the manufacturer's policing is strictly prohibited. But scrounging through plenty of reviews is a time consuming and demanding task and this project hopes to find the key by providing recommendations through analyzing structured and unstructured data which in turn would help manufacturers design their future products and upgrade existing ones to meet the needs of the consumers.

We can witness similar work has been done using product review data (Fang & Zhan, 2015), Opinion Mining and Sentiment Analysis (Pang & Lee, 2008), Product Weakness (Zhang et al., 2012) which lays the foundation of this project. Since this report aims at providing recommendations specifically about the SUV sub section of automobiles, it stands out and makes this project one of its kind.

The first step is by extracting data from automobile portals and then proceed with cleansing it. The cleansed data is now just a barrage of text which will be converted to sentences. Each sentence is now assigned a topic and then a sentiment is given to identify the emotion each line emits. The total positive comments are now divided by the total comments to figure out the appeal a particular component has on the general population.

Keywords: Data Extraction, Sentiment Analysis, StanfordNLP, Polarity Score

Contents

Chapter 1: Introduction

Just about everybody goes online before getting inline to buy something. Every customer want to hear what other people are saying about a latest product or service on sites like Zomato for food, Google for a destination or even Practo for how well a doctor consultations are. But the issue we discover here is companies work hard to get your trust through manufactured reviews and also aggregators make sure to filter out negative opinions therefore making the process of hand-picking a product grueling. Since one extra star on review sites increase the business by 9% (Luca, 2011).

It is evident that companies might be tempted to look at online reviews but how easy is it? As simple as loading a dataset into an algorithm and the outcome of this exercise would show a clear picture of which unit should the manufacturer work upon.

Sentiment Analysis/Opinion Analysis is a type of Natural Language Processing for identifying the emotions of the general population about a topic. This involves constructing a system to gather comments, reviews and opinions about the product made in a forum or a portal. Sentiment Analysis can be helpful in several ways like assisting in ad campaign of a product, figure out which product features are popular and even figure out the likes/dislikes of a particular feature.

Systematically listening to the perceptions, needs, wishes and expectations of your end users help the bottom line of the company's revenue. In July 2019, vehicle sales in India were 18.25 lakh units, compared to 22.45 lakh in 2018. The 18.71% fall was the steepest in 19 years forcing many manufacturers to cut production leading to the trimming of 2 lakh+ jobs in the automotive sector (Correspondant, 2019).

This unforeseen circumstance asked manufacturers to look at how they proceed with their past sales strategy. The customers are reluctant to buy which took a hit to the manufacturer's revenue. This project aims to solve the most important problem for the automotive industry as it has to rethink the way business is done by providing a solid strategy by assisting the manufacturers to keenly observe the sentiment of the customers and then re-working on the products to be released tailoring their needs. Some customers might be looking at buying BS6 vehicles which might have halted the income for the manufacturers to comply on the new emission standard. The industry giants are also worried that the government's focus on electric vehicles might postpone an end user's plans on purchasing a petrol/diesel vehicles.

The idea of this study would give us a list of most talked about features that customers liked/disliked about an SUV. The study concentrates on 8 SUV that is:

- Maruti Suzuki Vitara Brezza
- Hyundai Creta
- Toyota Fortuner
- Tata Harrier
- Tata Nexon
- Kia Seltos
- Mahindra Thar
- Hyundai Venue

Hyundai Creta – Figure 1.1

The reason these 8 SUVs were particularly picked was that it covered most of the companies that manufactured SUV in the Indian market. The project will now categorize the components of each SUV into categories that way the results are much more understandable and insights analyzed would be much more powerful. The topic modeling is built on Excel since LDA topic modeling, while it may be fast, requires lots of fine tuning, unable to influence topics and needs human intervention to label them to present the results to non-experts.

Considering these pressing situations, it becomes imperative that we observe and listen to our customer demands and introduce products accordingly. The Indian Auto Industry can only rebound if we are able to satisfy customer needs be it product features or improving customer service or both. The industry should delve into fresh and unique strategies of undertaking business which might be a new norm in future. With major home grown automobile manufacturer struggling to regain their sales, maybe it is due we consider customer reviews before we implement unsubstantiated strategies during production.

Chapter 2: Literature Review

An archaic solution to getting customer reviews would be through telephonic or face to face surveys and even that would not guarantee you an undressed emotion an audience is carrying unto themselves. It is much more effortless to get your information through forums since everyone has access to cell phones and internet is now a commonplace (Kuhn, 2017).

It only make sense to be on the same wavelength with technology. Identifying the voice from the customer requires scraping data from various blogs, forums and tweets available through social media and this technique is known as text mining.

Text Mining is used to help us answer specific research questions. It would be impossible for a human to read millions of research articles manually on the topic picked and this is where Text Mining can help us immensely.

According to a study in (Salloum et al., 2018) text analytics generally refers to a process of extracting the interested information from a unstructured text which is further used for word exploration to study frequency spread, pattern identification, tagging or annotation, taking out information, data mining methods including link and association analysis, visualization and predictive analytics (Ranjan et al., 2015).

A key element about text mining is the linking together of the extracted information together by mapping pattern and trends across millions of articles, reviews and comments (Hearst, 2003).

This detailed relevant information helps us determine what additional research is neede to answer our question so now we can go back into the lab with a headstart in order to do further research.

Recently Text Mining has quickly evolved by adding methods able to classify documents according to their latent topic or to infer about the "sentiment" of customers or the users of social networks. The push to these techniques is parallel with the evolution of both the computational efficiency of the algorithms necessary to analyze text data and the technology needed to capture information (Zappa et al., 2019).

A lot of work has been done on Sentiment analysis using product review data (Fang & Zhan, 2015), Opinion Mining and Sentiment Analysis (Pang & Lee, 2008), Product weakness finder (Zhang et al., 2012), Sentiment Analysis: A Multi-Faceted Problem (Liu, 2010a), Comparative Experiments on Sentiment Classification for Online Product Reviews (Cui et al., 2006), which form the basis of our work.

However, Sentiment Analysis remains a very challenging proposition since our knowledge of the problem and solution are very limited as it deals with Natural Language Processing (NLP). Also relying too much on Machine Learning Algorithm produce no human understandable results (Liu, 2010b).

Since we have used the comment section for detecting sentiment and using lexical resources to capture information about the informal/casual language used in web forums, usage of sentiment lexicon proved useful to determine positive/negative/neutral sentiment (Kouloumpis et al., 2011).

We have performed TextBlob by studying the existing frequency of the word and how often they appear in a positive/negative/neutral context and decide the polarity (Singh et al., 2017).

This project focuses on Indian Auto industry on SUV sub section, one of the fastest growing market in the world, contributing to a large share in Indian economy (Verma, 2019). India is poised to be world’s third-largest passenger-vehicle market by 2021. Currently, the automotive sector contributes upwards of 7% to India’s GDP. Mini and hatchbacks have found a commonplace for the automobile field in India, with marketshare around 50% and growth of 6-7% between fiscal year 2014 and 2017. Minis and Hatchbacks will continue to maintain a superior stand, but the majority of growth is expected to arrive from new segments such as compact SUVs, sedans, and luxury vehicles. In the Automotive Mission Plan 2026, the government along with the industry fixed a goal to triple revenues to $300 billion, and expand exports 7x, to $80 billion. To reach the targets, it is evaluated that the sector could come up with more than $60 million additional in/direct jobs and the outcome could be improved manufacturing competitiveness and reduced emissions (BMR Advisor, 2015).

As a result many new entrants are vying for a slice in such a huge market. This will increase competition, reduce product lifecycle and give customer more choice. Now to move and adapt fast and gain upper hand in this dynamic and challenging market, it is very important how well an industry understands and responds to customer needs. This is where our study can be of immense importance and it seeks to identify the features that customers are expecting or not expecting by analyzing the structured and unstructured text data available in Team BHP/Overdrive portal, the first Indian automotive community to have cemented its position on the Global Big Boards Registry, the most comprehensive and only unbiased information resource on the Indian automotive scene where car owners & enthusiasts come together to form a collective force providing honest and unbiased comments (Team BHP, n.d.).

This would help car manufacturers design their future product better and upgrade existing products to meet the needs of customers. Providing recommendations to the car industry based on insights extracted from mining web portal will make our project one of its kind as car industry has so far not been covered under any of the available studies.

Chapter 3: Problem Statement

The study done in this project focuses on comparative metrics of 8 selected SUVs on the top 6 most talked about feature. It shows the contrast of emotions of general population between the SUVs and the components and would help manufacturers take a look at why their components might be invoking a hostile sentiment and work upon it.

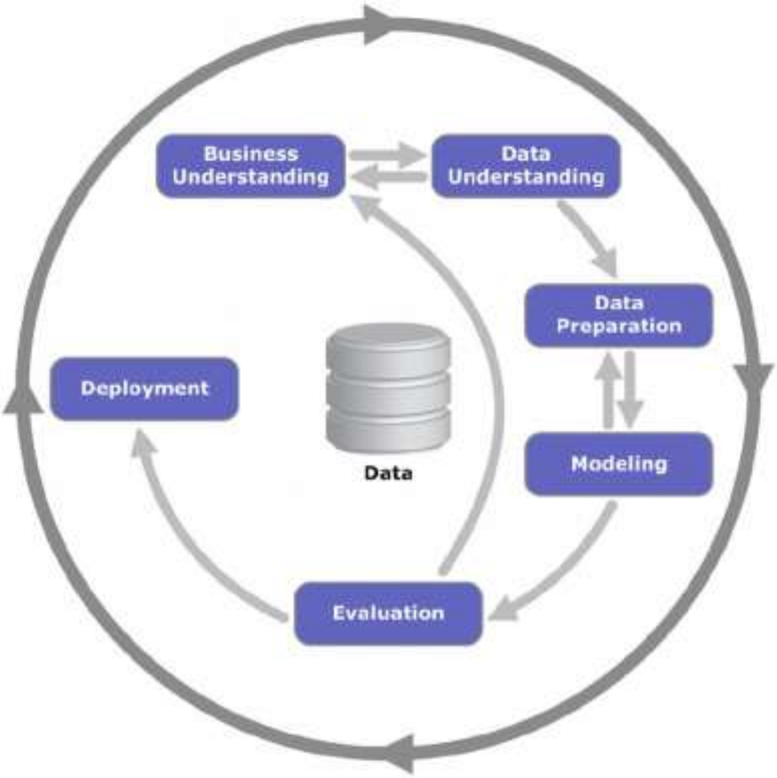
Systematically listening to the perceptions, needs, wishes and expectations of your end users help the bottom line of the company’s revenue. This would help SUV manufacturers design their future product better and upgrade existing products to meet the needs of customers.

Chapter 4: Objectives of the Study

The aim of this project is 2 fold:

- What feature is most talked about in an SUV and whether those are positive, negative or neutral?
- Compare SUV's components (topic) with other entries in the list and look at how each of them are performing against each other.

The study uses the CRISP – DM Data Science Cycle.

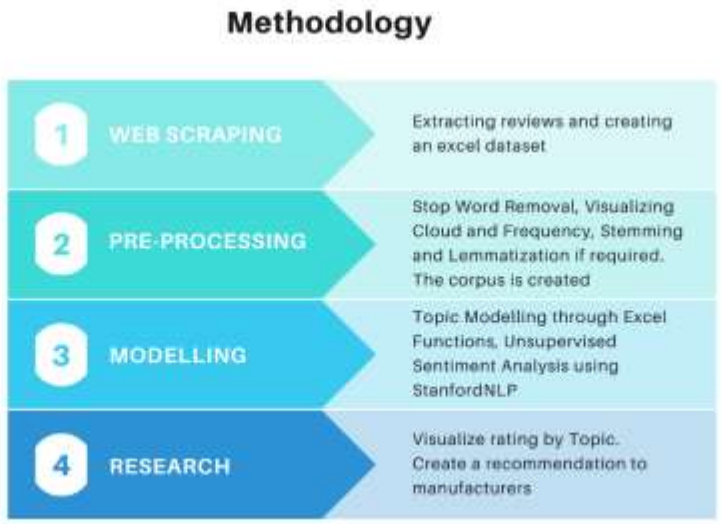


CRISP – DM Cycle – Figure 4.1

The project aims to tackle these requirements by analysing data available in portal like Team-BHP. All these put together will become our recommendation model for the SUV Auto Industry.

As the auto industry contributes approx. 7% of the country's GDP and it is said to triple the revenues by 2026. The claim can be validated by looking at the current demand for SUV segment cars (Mehra, 2020),(Nadkarni, 2020).

Chapter 5: Project Methodology



Methodology

1 WEB SCRAPING — Extracting reviews and creating an excel dataset

2 PRE-PROCESSING — Stop Word Removal, Visualizing Cloud and Frequency, Stemming and Lemmatization if required. The corpus is created

3 MODELLING — Topic Modelling through Excel Functions, Unsupervised Sentiment Analysis using StanfordNLP

4 RESEARCH — Visualize rating by Topic. Create a recommendation to manufacturers

Project Methodology – Figure 5.1

This project is a product development technique that is used to collect feedback from consumers. Since customer experience is the major differentiating factor against rival companies or competitors. We collect the data from a relevant website and do analysis on the product’s review and get back with a sentiment analysis with respect to the product.

- Web Scraping – Comments displayed by most websites can only be seen using a web browser like Chrome/Firefox/IE and as they do not offer a functionality to save the copy of the data in the local drive for personal use, one of the option is to manually copy paste data into an excel sheet which turns out to be a tedious job and can take many hours depending on how large the dataset is recommended to perform the analysis. Web Scraping is a method to automate this tedious process. So instead of manually performing the task, a web scraping software can be used to do the same task in a fraction of time.

  ParseHub is a web scraping software which is used to pull data from the forum/portal. The comments are extracted using this software where the URL to the comment section is inputted. Now the inspect element of the browser page is opened and the comment/reply body is selected, that way the software recognizes which data is to be scraped through its algorithm. The below image talks about Maruti Suzuki Vitara Brezza and as stated, the thread has 113 pages. The data is scraped for all 113 pages including comments and replies to the comments and is extracted into a csv file.

- Data Pre-Processing – This method involves transforming raw data into a useful or efficient format. With the raw data scraped from the TeamBHP portal, the dataset now consists of 13470 comments for all the SUV selected for the analysis. Pre Processing involves cleansing the data since noisy data cannot be interpreted by machine learning algorithm which will used down the lane in this project. This is handled by removing stop words, Visualizing Word Cloud and Word Frequency to figure out what is the most commonly used words. The most commonly used words forms our topic in the Modelling section which will be discussed in the next paragraph.

- Modelling – The most common words are now filtered to look for feature specification. As we discussed in the previous paragraph about allocating topics, the most used words which talk about features are hand pick to be the talking point of a particular comment. Each review is now assigned a topic i.e. category that way it is easier to understand which particular feature a customer likes/dislikes. The data is run through the StanfordNLP algorithm on Python which gives the score of positive, negative or neutral for each review. Since StanfordNLP does not require the use of train data, it is easier to figure out the emotion each review is emitting without pre training the model for accurate results. Once each comment is analysed for sentiment, a bar chart is now plotted for each SUV and each feature specification individually to better understand which specs are most liked/disliked by the customer. Another bar chart is plotted for comparison between SUV and their features to realize which SUV is performing better among the customers.

- Research – The visualization helps to build insights on the current emotion among the SUV owners in the general population. The research done would suggest manufacturers what component/feature to be altered in the upcoming iteration on the SUV and the insight would even help manufacturers in the decision making process of building new SUV without leaving out crucial components which would upset the customers and end up being a disappointed product.

Chapter 6: Business Understanding

The auto sector industry is a key driver of the economy frankly across the globe with a huge multiplier effect and in India it is no different. The country’s automotive sector constitutes 7% of the GDP and 49% of manufacturing GDP (Bodke, 2019). Inclusive of its value chain, it supports 3.7 crore jobs. There are a few things that can be done to move together that can benefits the manufacturer as well as the consumer. On the resource side, the company needs support on looking at what the end user are asking for and on the consumption side, they need to know whether the features currently existing on a product satisfies their needs. This project has the right steps being taken that will lead to a better relationship between the company and the consumer. The course of action necessary for the revival has been put in place and the future will look brighter if adopted.

Since the government’s measures to boost the economy are genuine and far sighted, there are steps to be taken from the manufacturer’s side as to not be completely dependent. The cost to be incurred while making changes to the existing product depends on each company but it is better to know what changes has to be made rather than speculate the likes of a consumer and then release a detested product.

Currently India is expected to turn out the world’s 3rd largest passenger vehicle by 2021 and is expected to reach $300 Billion by 2026 (Bajwa, 2019).

Since post Covid, we saw the auto industry chart stepping towards the greener pastures and with the market share being large enough, manufacturers will be looking a slice of that pie. The strategy to be used will be discussed further.

Chapter 7: Data Understanding

Data Understanding is a crucial step in the CRISP – DM data science cycle. By working on a clear Data Understanding, the project comes closer to an agreeable success criterion.

Some websites can contain a large amount of valuable data, be it stock prices or reviews in form of comments in our case. If there is any chance to use the information, it will require the code to be built for each page in a different way which ends up being tedious and also websites do not typically tend to allow web scraping as it creates tremendous load for their database and servers and might end up crashing the websites. The move also lies on a grey area in the eyes of the law too and the website tends to block the IP of users found scraping data. Although web scraping can be done manually, there are software tools that exist for this very reason limiting the time used to pull data manually and automating it end to end. But since a lot of websites are designed differently, web scraping can differ a lot too when done through automation.

ParseHub is used to pull data from websites since it masks the IP address for every user and scraps the data without creating a load the website server creating a win – win situation for both the website and the user. Each SUV official discussion page is now inputted one by one in the software and then the whole thread including comments and replies and asked to scrap from the website. This is done by mentioning the total page number count for each SUVs thread.

SUV Name	**Total Pages ; Total Reviews Extracted**
Maruti Suzuki Vitara Brezza	113 ; 1628
Hyundai Creta	170 ; 2390
Toyota Fortuner	51 ; 1166
Tata Harrier	72 ; 1556
Tata Nexon	223 ; 2430
Kia Seltos	95 ; 1546
Mahindra Thar	153 ; 2430
Hyundai Venue	22 ; 324
Total	**899 ; 13470**

Reviews and Pages – Table 7.1

The above table states the Total Pages and the Total Reviews extracted for each SUV. Data Understanding asks the question whether the data extracted for this project relevant to the problem the project is trying to solve. The answer is a simple yes, as the data pulled from automobile review forums specifically filtering for SUV sub section of the industry.

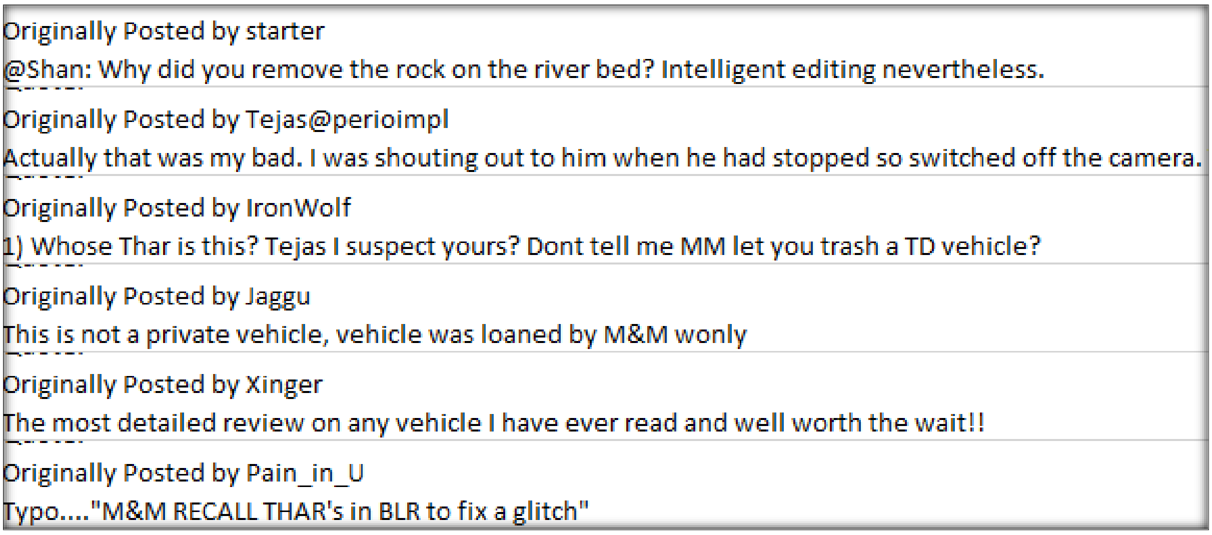
Chapter 8: Data Preparation

Data Preparation is one of the most tedious step in the data science project life cycle since data is a key tool in constructing a model. This process provides a context in which we can consider the data preparation required for the project. It exposes the unknown underlying structure of the problem to algorithms.

Since the data is scraped from various automobile forums with thousands of comments making up as our dataset, before loading the data into the model, make sure it is cleansed and parsed that way it can be readable by the machine learning algorithm StanfordNLP. The below points reflect clearly on what necessary steps are to be taken.

- Manual Intervention – The data is to be glanced through manually to catch any abnormalities existing. The anomalies that can be witnessed here are unnecessary use of the word ‘Originally Posted’. The phrase exists since a thread contains discussions among the general population and users tend to reply to the original comment made furthering the discussion even more. Every reply to an original poster contains that phrase which has to be eliminated in order to not decontaminate the corpus. The method used here would a straightforward ‘Find and Replace’. Find “Originally Posted”; Replace with “”. This would automatically find every phrase in the corpus and then replace it with a blank.

  Before



Originally Posted by starter
@Shan: Why did you remove the rock on the river bed? Intelligent editing nevertheless.

Originally Posted by Tejas@perioimpl
Actually that was my bad. I was shouting out to him when he had stopped so switched off the camera.

Originally Posted by IronWolf
1) Whose Thar is this? Tejas I suspect yours? Dont tell me MM let you trash a TD vehicle?

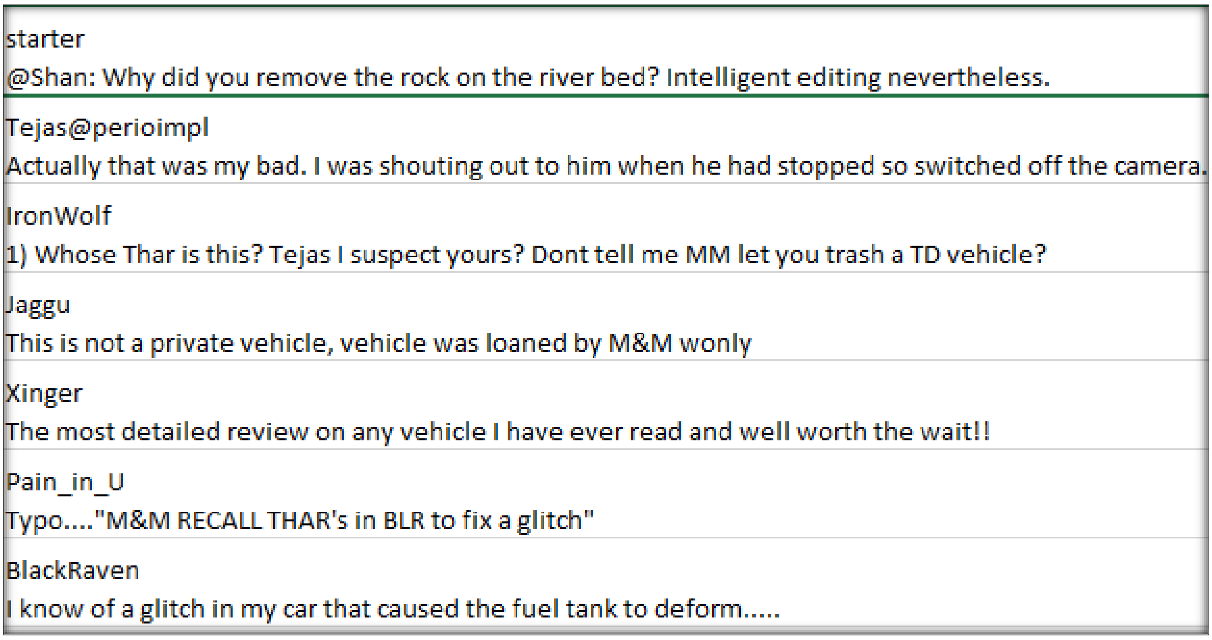
Originally Posted by Jaggu
This is not a private vehicle, vehicle was loaned by M&M wonly

Originally Posted by Xinger
The most detailed review on any vehicle I have ever read and well worth the wait!!

Originally Posted by Pain_in_U
Typo...."M&M RECALL THAR's in BLR to fix a glitch"

Before StopWords – Figure 8.1

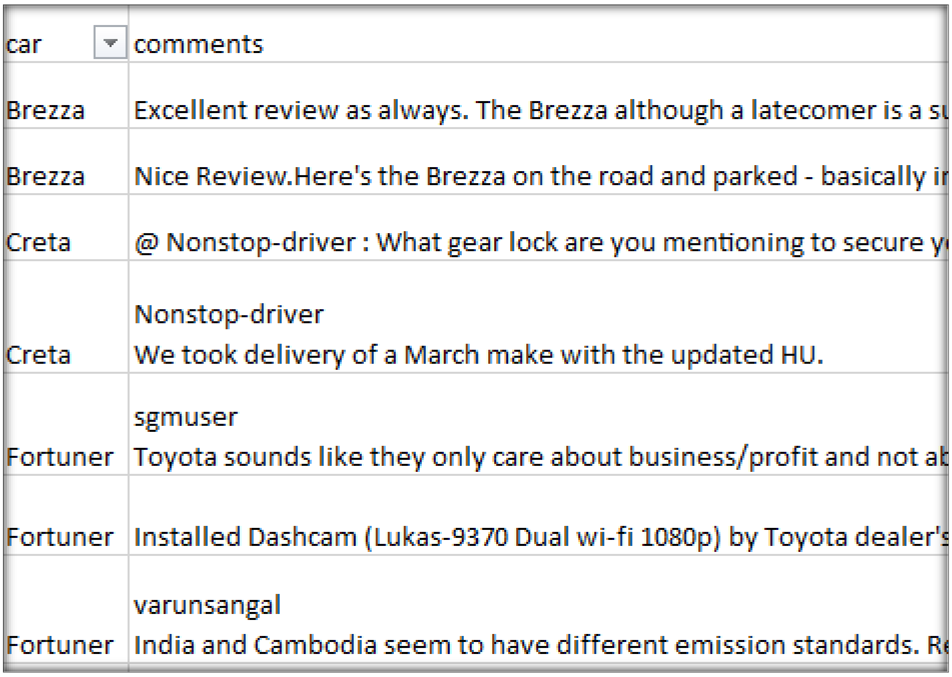
After



starter @Shan: Why did you remove the rock on the river bed? Intelligent editing nevertheless.
Tejas@perioimpl Actually that was my bad. I was shouting out to him when he had stopped so switched off the camera.
IronWolf 1) Whose Thar is this? Tejas I suspect yours? Dont tell me MM let you trash a TD vehicle?
Jaggu This is not a private vehicle, vehicle was loaned by M&M wonly
Xinger The most detailed review on any vehicle I have ever read and well worth the wait!!
Pain_in_U Typo...."M&M RECALL THAR's in BLR to fix a glitch"
BlackRaven I know of a glitch in my car that caused the fuel tank to deform.....

After StopWords – Figure 8.2

- Append – Since the data is collected individually for every SUV, the output received also exists individually for every SUV, meaning, the data for every SUV is stored in separate csv files. This step involves appending all the existing csv files into 1 single csv file to be loaded into the model so as to not run the algorithm multiple times.



car	comments
Brezza	Excellent review as always. The Brezza although a latecomer is a su
Brezza	Nice Review.Here's the Brezza on the road and parked - basically ir
Creta	@ Nonstop-driver : What gear lock are you mentioning to secure y
Creta	Nonstop-driver We took delivery of a March make with the updated HU.
Fortuner	sgmuser Toyota sounds like they only care about business/profit and not ab
Fortuner	Installed Dashcam (Lukas-9370 Dual wi-fi 1080p) by Toyota dealer's
Fortuner	varunsangal India and Cambodia seem to have different emission standards. Re

Post Appending – Figure 8.3

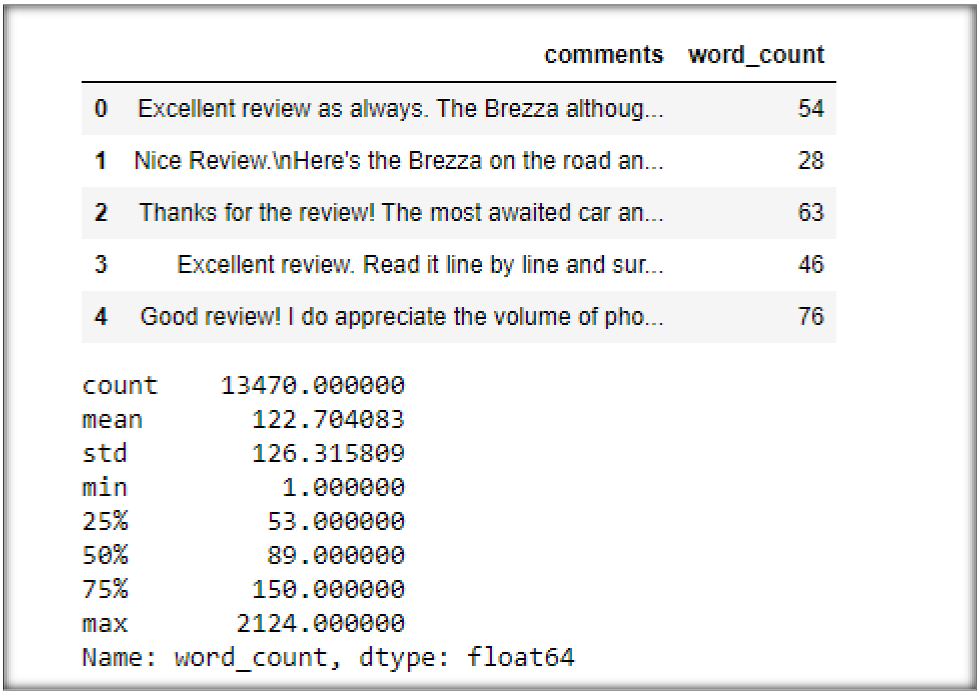
- Arranging – The data is now brought into a single file as mentioned in the previous step but is yet to be categorized as per the SUV that the EDA performed later down the project is easy to gather insights. The data is assigned its respective SUV. The above image in the ‘Append’ step shows about the comments appended with SUVs in the first column.

The data is now ready for the next step which is Modelling.

Chapter 9: Data Modeling

The previous step in the data science cycle was to prepare the data for modelling purpose. All the data is cleaned and ready to be fitted in the algorithm. Descriptive Statistics/Exploratory Data Analysis needs to be run on the model to get the results on how the data looks. Since this is just a barrage of text, looking at how frequently a word is used by building a word cloud, a visual representation of the words that occur most often, words that stand out, words that are used a lot in a given dataset will help us prepare for the topic/feature for the model. Word frequency is built on the dataset which sheds light on the number of times the most common words are used. This procedure helps us identify redundant data to that needs to be sanitized. Please note, stop word removal is only to be used to identify the most talked about feature. The corpus is again reinstated with those same stop words. The reason is explained later down in the same step.

Descriptive Statistics – Superficial Descriptive Statistics is to thoroughly understand the data. It is needed to unearth the elementary structure of the data and is important since it reveals trends, patterns and relations that are not easily known. It is unfeasible to draw well grounded conclusions from a huge quantity of data by just glancing over it, instead it is recommended to have a look at it carefully through an analytical lens. This critical information can help detect mistakes.

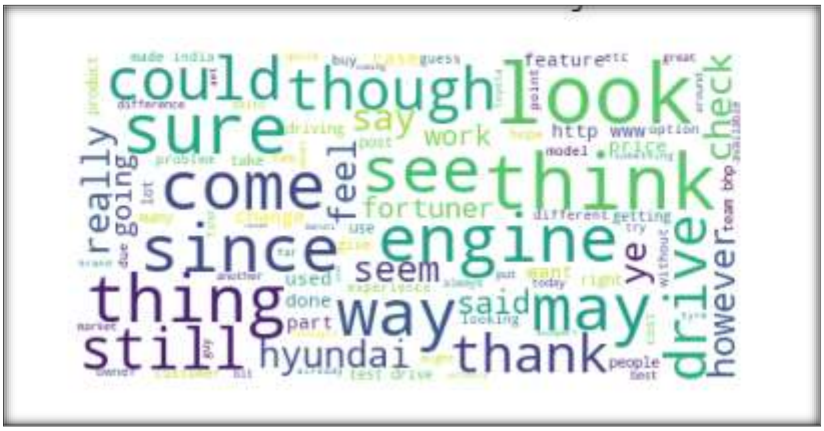


	comments	word_count
0	Excellent review as always. The Brezza althoug...	54
1	Nice Review.\nHere's the Brezza on the road an...	28
2	Thanks for the review! The most awaited car an...	63
3	Excellent review. Read it line by line and sur...	46
4	Good review! I do appreciate the volume of pho...	76

```
count    13470.000000
mean       122.704083
std        126.315809
min          1.000000
25%         53.000000
50%         89.000000
75%        150.000000
max       2124.000000
Name: word_count, dtype: float64
```

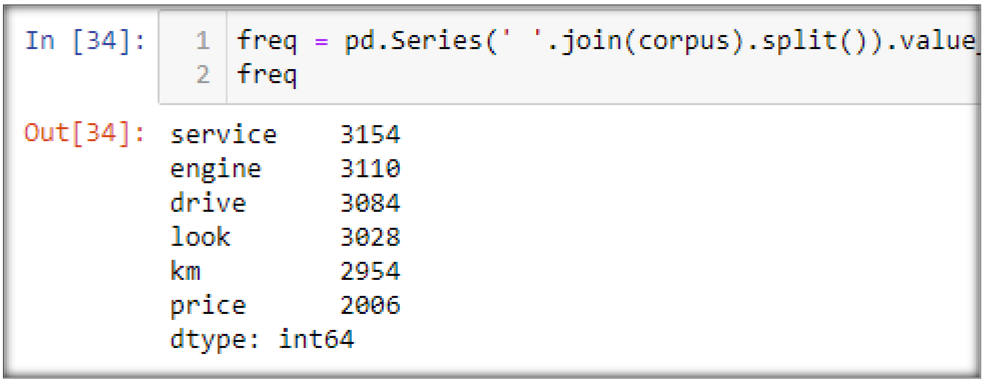
Descriptive Statistics – Figure 9.1

Word Cloud – This is a visual representation of the most frequently used word. Since Word Frequency already does a better job than word cloud for a much concise results, word cloud helps understand visually.



Word Cloud – Figure 9.2

Word Frequency – This process is done to figure out the most commonly used feature/specs in the corpus. Stop Word is usually done through NLTK library in python which will be used in the modelling process. Additional custom Stop Words are also added since NLTK Stop Words only tend to eliminate common words like ‘the’, ‘and’, ‘is’ etc. Getting rid of additional custom Stop Words helps identify the topics as shown below.



```
In [34]:  freq = pd.Series(' '.join(corpus).split()).value_
          freq

Out[34]:  service    3154
          engine     3110
          drive      3084
          look       3028
          km         2954
          price      2006
          dtype: int64
```

Word Frequency – Figure 9.3

The most frequent word in the above image is now to be used as the major topics to be discussed in the model. 6 Topics are considered for analysis.

Topics	Definition	No. of Comments
Service	The overall customer service	764
Engine	The performance of the engine	1476
Drive	How well it performs overall	716
Look	The aesthetics of the car	1517
KM	Mileage	748
Price	Is it worth the value?	3731
Other	This topic is added since not all reviews will fall into the above 6 categories. This topic will not be included in the analysis	4518

Topic Selection and Definition – Table 9.1

Every review is now assigned a topic. This is mostly done on Microsoft Excel environment since there would a lot of control on what keywords are assigned to the topic. The corpus now only contains 8952 relevant comments since reviews falling under 'Other' are eliminated from the analysis.

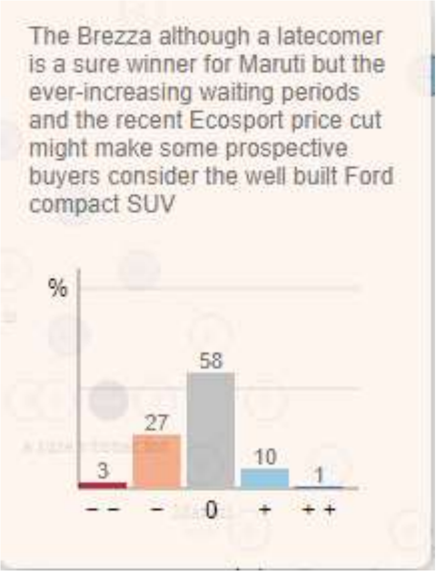
Introducing StanfordNLP - The dataset now contains multiple variables with the review making up many of that since each review is split into sentence and then its respective topic. This is proceeded with expanding the variables by adding sentiment to each sentence from a particular review by using the StanfordNLP Library in Python. StanfordNLP is a new Natural Language Processing library available in Python. If you are familiar with NLP, there are a lot of library which can be used in Python like NLTK and Spacy but StanfordNLP stands out in some areas than the rest, being other libraries highly rely on hand crafted rules and patterns, for instance, variety of libraries have to be installed for tokenization or lemmatization. StanfordNLP has fully data driven modules for easy domain adaption and pre-trained neural models supporting 50+ human languages. This means we are looking at one of the most robust NLP libraries Python has to offer which do not require us to train a model to get good results.

Deciding to reinstate the stopwords back into the corpus because StanfordNLP reads the entire sentence and then produces a score. Removing the stopwords would remove crucial information that would require the algorithm to work seamlessly. Proceed with using

Dependency Parser which takes a sentence as inputs and learns to produce a tree structure that represents the syntactic structure of a sentence. This enables the model to get the sentiment of each sentence into 3 scores .i.e Positive, Negative and Neutral. The scores a now assigned as number 2, 1 and 0 respectively that way the calculation of percentages is simpler. Once the assigning is done, aggregating each topic based on their scores and figuring out the Positive, Negative and Neutral score for each topic and for each SUV car is carried out. From here the user gets to know how each SUV is performing in every component listed in the topics of its product. Understanding the emotions and the sentiment being emitted from the audience about their unit is now a reality.

The below images shows how the StanfordNLP works

Neutral Review -



The Brezza although a latecomer is a sure winner for Maruti but the ever-increasing waiting periods and the recent Ecosport price cut might make some prospective buyers consider the well built Ford compact SUV

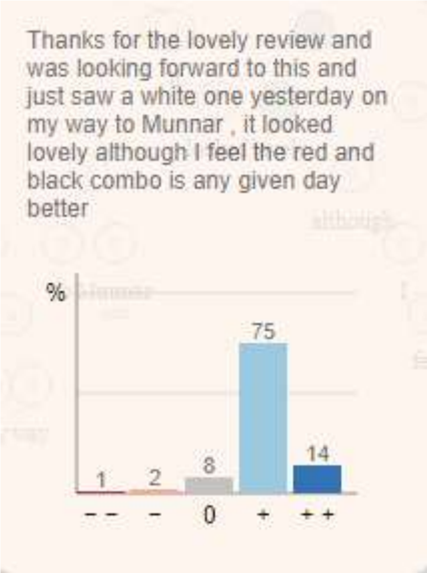
Neutral Score – Figure 9.4

Each review is given a total score of 100 and a polarity score from -2 to +2.

The comment above talks about Brezza being a winner (positive) yet also talks about how the Ford EcoSport might cannibalize the market share (negative). Since both polarity come at play, StanfordNLP declares this review as Neutral.

Positive Review -

Let us take a look at a positive comment.

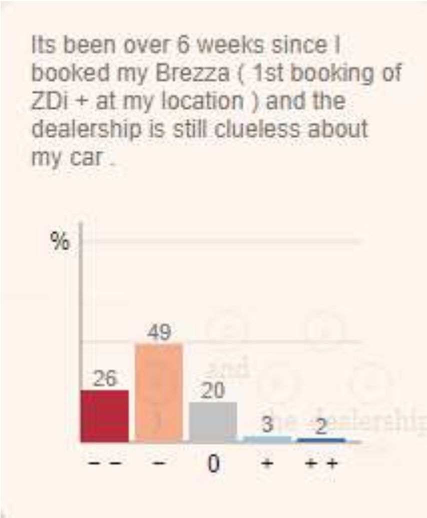


Thanks for the lovely review and was looking forward to this and just saw a white one yesterday on my way to Munnar , it looked lovely although I feel the red and black combo is any given day better

Postive Score – Figure 9.5

The reviewer clearly states how they were looking forward to the SUV and likes the look of the car. 75 of 100 is assigned a score of +1 and 14 of 100 is assigned +2.

Negative Review –



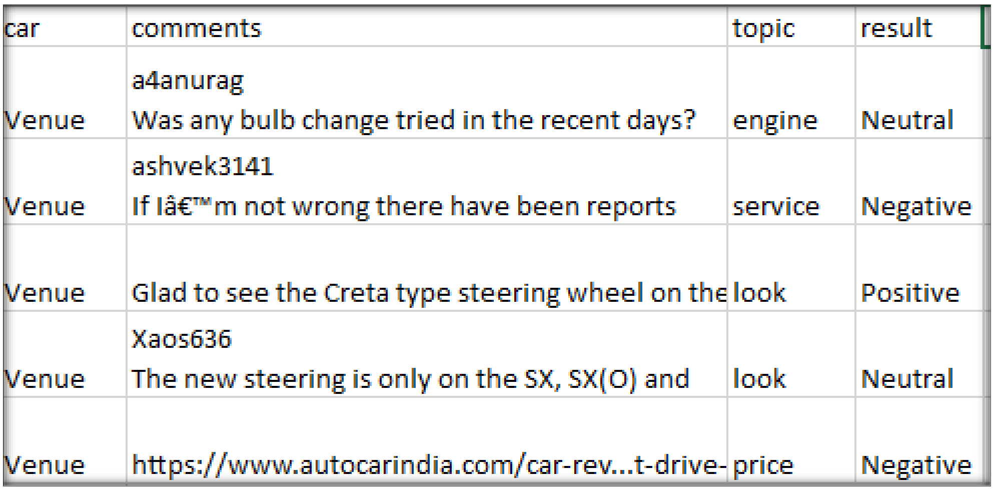
Its been over 6 weeks since I booked my Brezza (1st booking of ZDi + at my location) and the dealership is still clueless about my car .

Negative Score – 9.6

The above reviewer seems upset about the service provided which made the model give the review a score of Negative.

Chapter 10: Data Evaluation

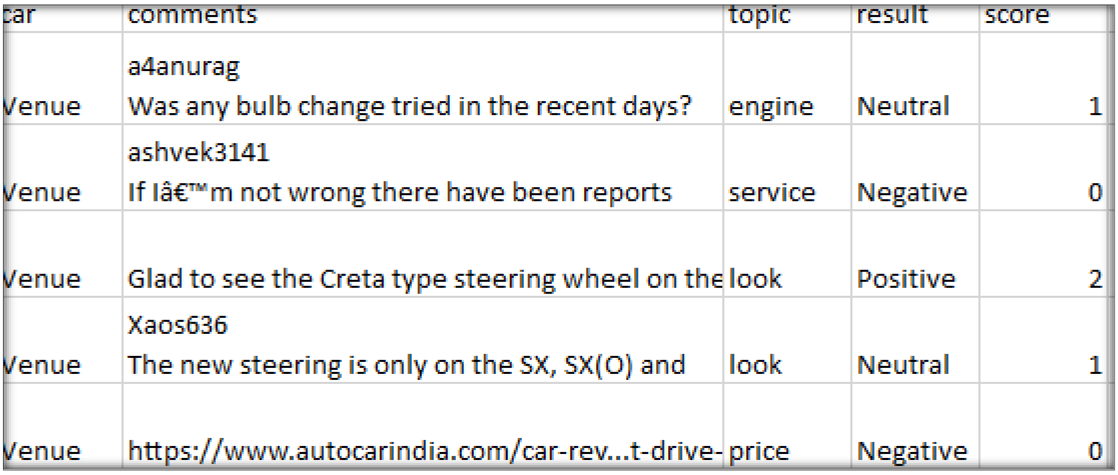
Since the model is now complete, let us take a look the evaluation on how the model fits our business understanding. The business understanding was to boost the sales of the Indian Auto Industry after it took a huge slump in the market and what measures were to be taken during the period without affecting the hole in an already sinking boat. The results reflect on what component the audience are showing positive response and where the opportunities lie for manufacturers for them to increase their concentration on. The concentration should lie on how the results are helpful to the business by assessing the model on the results it has given and rework if any changes are to be made or if the model does not fit the business objective. This model fits the objective and proceed with the results.



car	comments	topic	result
Venue	a4anurag Was any bulb change tried in the recent days?	engine	Neutral
Venue	ashvek3141 If Iâ€™m not wrong there have been reports	service	Negative
Venue	Glad to see the Creta type steering wheel on the	look	Positive
Venue	Xaos636 The new steering is only on the SX, SX(O) and	look	Neutral
Venue	https://www.autocarindia.com/car-rev...t-drive-	price	Negative

Result – Figure 10.1

The ‘result’ column contains the sentiment score for each review. Another column is now added as a numeric version to the result column which makes it easier to calculate the percentage and compare SUVs with each other.



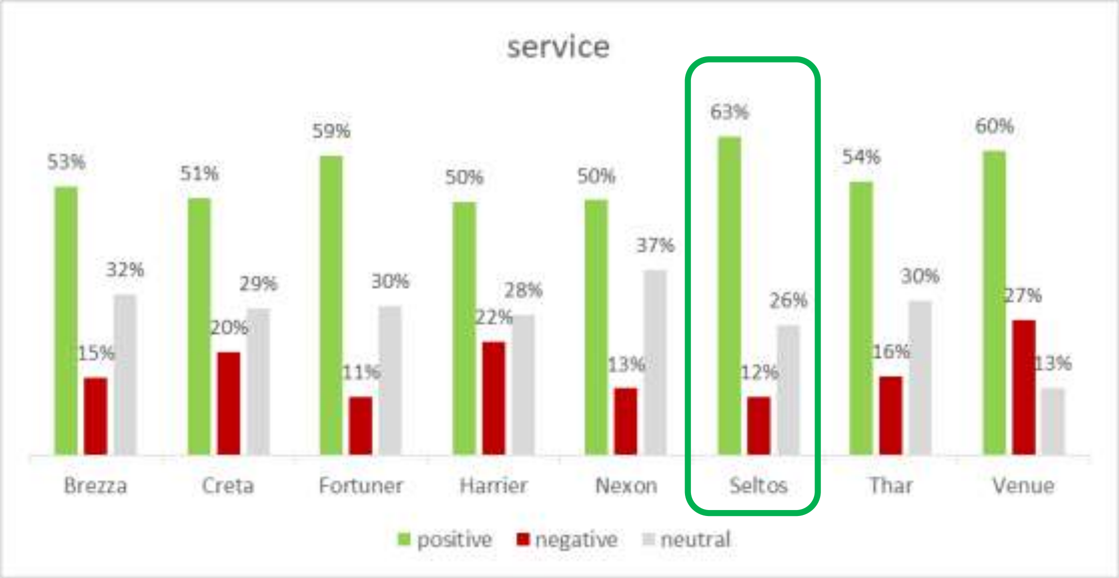
car	comments	topic	result	score
Venue	a4anurag Was any bulb change tried in the recent days?	engine	Neutral	1
Venue	ashvek3141 If Iâ€™m not wrong there have been reports	service	Negative	0
Venue	Glad to see the Creta type steering wheel on the	look	Positive	2
Venue	Xaos636 The new steering is only on the SX, SX(O) and	look	Neutral	1
Venue	https://www.autocarindia.com/car-rev...t-drive-	price	Negative	0

Score – Figure 10.2

Each score for an SUV/Specs is calculated with the total of its particular category.

Chapter 11: Analysis and Results

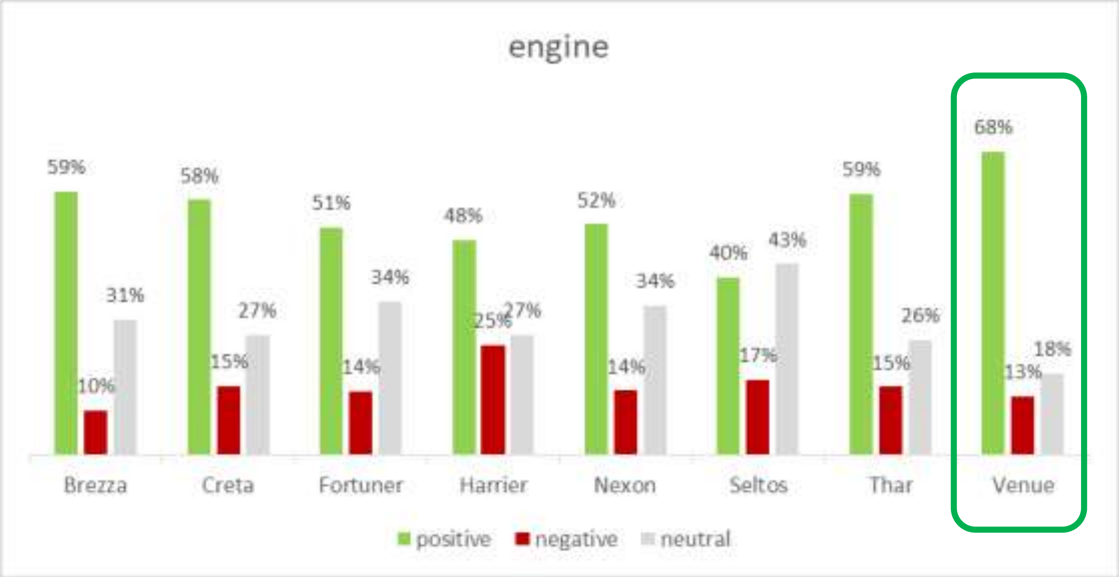
Analysis of each SUV and comparison charts for each component. To better understand which SUV is performing well in each category, we use the formula of Positive % - Negative % to balance the percentage out.



Service – Figure 11.1

Kia Seltos scores the highest among the SUVs in the ‘Service’ category with comments like “*Looks like there is no stopping on Seltos trend.*

Kerala's top Kia dealer has just delivered 100 Seltos in 4 days post lockdown relaxation!”



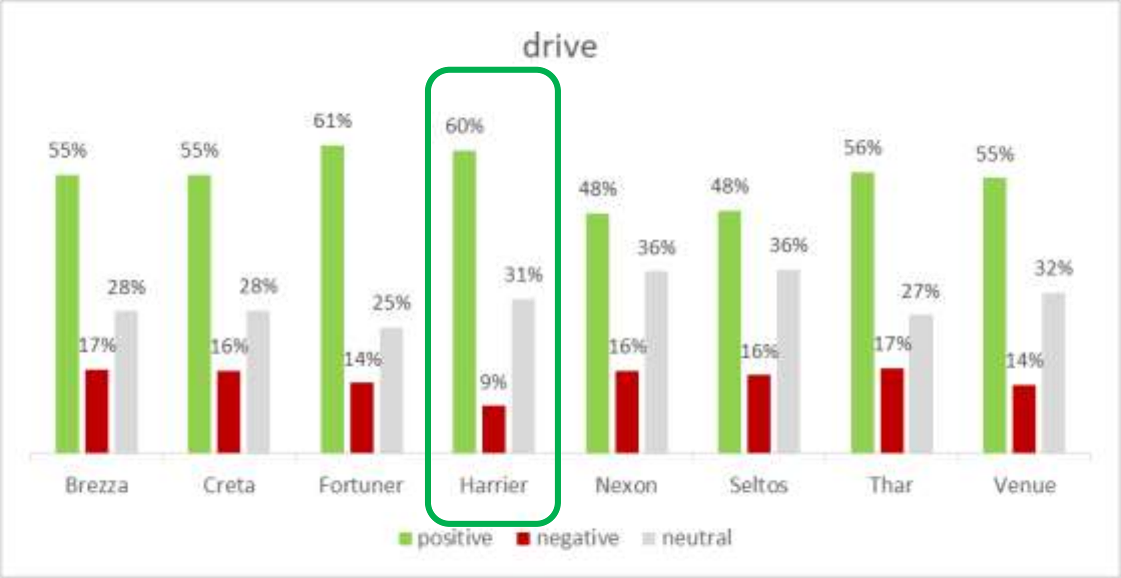
Engine – Figure 11.2

Venue is considered the best when it comes to engine performance. Let us take a look at what promit from TeamBHP has to say.

“*It was my first time driving a manual turbo petrol. Man, the thing goes!. It's hard to not want to drive like a hooligan. The engine sounds a bit gruff but loved the power delivery and response of the motor. I found the drivers seat comfortable. The bells and whistles (escpecially the sun roof) are nice to have. Interesting little package the Venue is, i must admit.*

Agree. I just took a test drive of the Turbo Petrol SXO Manual and its hard not to like this car. If you are OK ferrying 4 adults and a kid or maybe 5 adults occasionally, the car works really well.

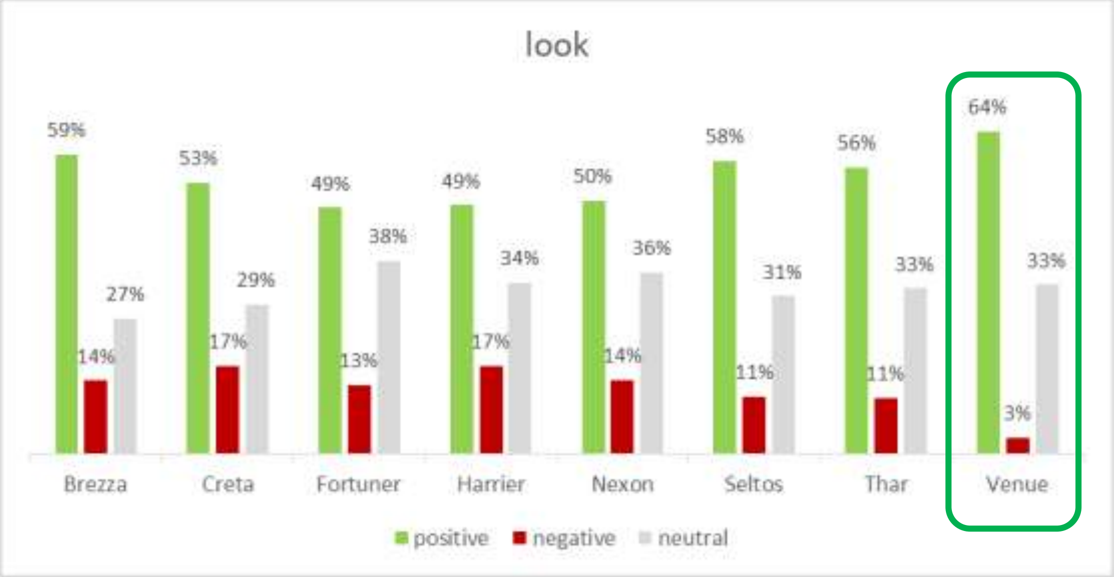
The Turbo petrol engine is quite awesome. It revvs really well and clean all the way to 6500rpm in all gears and the car flies. It is even better than the K series engine, coupled with the fact that it has 120 horses. You can drive the manual car like an automatic. It was able to pull from 600rpm with 3 adults on-board without any problem in 3rd gear. I might have driven it for some 25kms during the test drive. It is even more awesome to drive on the highway as I found out. The engine is not harsh, the NVH is terrific. The car is light and effortless to drive. It returned 14km/l during the test drive in mixed traffic conditions. It was not a B-2-B traffic and more of free flowing traffic.
Really impressed with this car and I would buy it just for the engine and gear box combo. I will check out the Brezza(when launched)”



Drive – Figure 11.3

Harrier takes the lead for the overall feel of the SUV and is evident when comments like the below float around the forums.

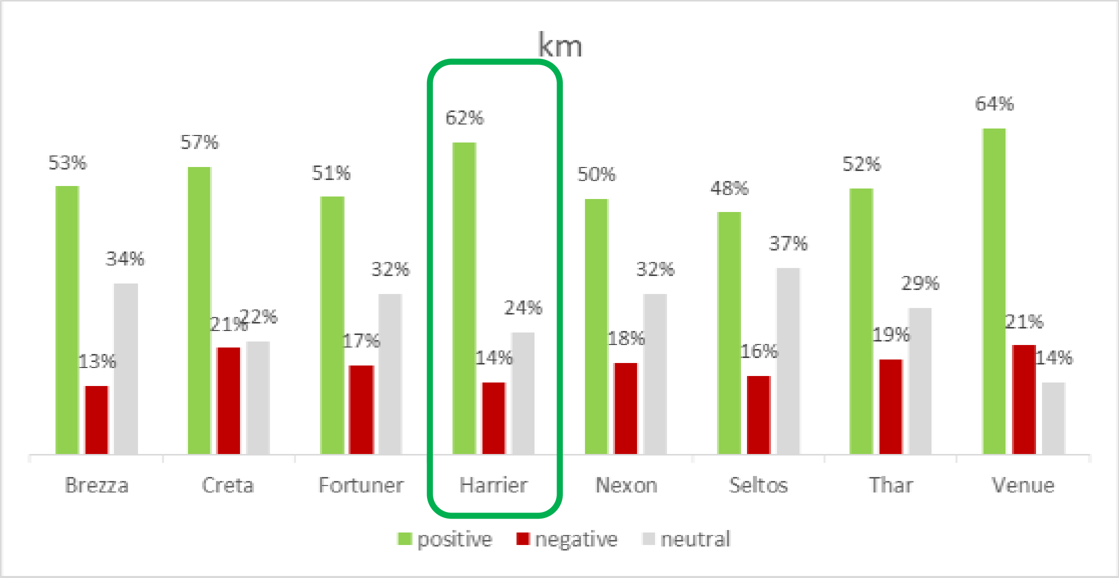
“*Harrier rides way better than the Hector, and is more stable on the highways, in my opinion. Have driven both and I'd pick the Harrier anyday. So basically, this has the same power figures as the compass and now it'll (hopefully) have a good, responsive autobox.*”



Look – Figure 11.4

Venue is considered the best for its aesthetics since it has a whopping 64% positive reviews and just 3% negative.

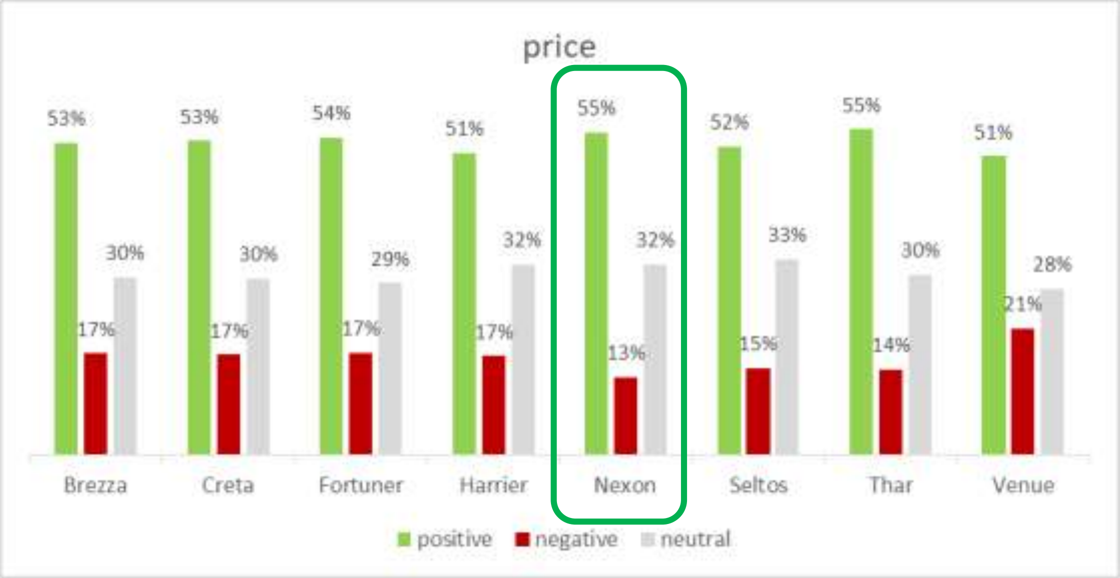
"Congratulations on your purchase, the colour looks dashing on your car"



KM – Figure 11.5

The general population tend to believe that Harrrier seems to have an edge over the mileage than the other SUVs competing.

"So on 17 July 2020, our Harrier XZ completed one year, and it's safe to say the journey has been very beautiful and pleasant till now. Slowly inching towards 10k kilometres (would've certainly done more if not for this pandemic, done only about 500kms since March). City mileage is around 11 to 13KMPL while highway mileage is about 16 to 18KMPL. The best I've attained was 20KMPL while I was really hunting for the best mileage figures"



Price – Figure 11.6

Nexon are considered the best value for their price with their relatively low price of 9 Lakhs for a full loaded.

This concludes the Analysis and it is clear which SUV has an edge over the other entries in the given categories but the manufacturers can look into what the component the general population are not preferring/dislikes and work on the weakness on their SUV.

Chapter 12: Conclusions and Recommendations for future work

It is now known how well the algorithm is performing against the dataset. When looked at an SUV's component which has received certain negative traits, the company can rethink their strategy and work upon improving the existing issue on their upcoming iteration.

The engines of India's economy have slowed down. The auto industry says this is the worst crisis in nearly 2 decades which is taking a toll on the country's GDP. Factories are reducing work hours to cut costs which means less pay. This study recommendations clearly states the metrics an SUV needs to fulfill in order to meet the customer demands. While there are other factors which altered the course of the auto industry's decline in this nation, the primary focus always remain customer satisfaction. While the study focused on SUV category of automobiles, further study of this will require are to include Hatchbacks, Mini SUVs and Sedan.

The plan also includes to incorporate multiple categories capturing the nitty gritty of the car components that way no page is left unturned.

Bibliography

Bajwa, N. (2019). No Title. *Investindia*.

BMR Advisor. (2015). Indian automotive industry: The road ahead. *Forbes India*.

Bodke, A. (2019). Trouble in Motown! Auto sector which is 49% of manufacturing GDP looks for rescue. *MoneyControl*.

Correspondant, S. (2019). Auto sales fall 19% in July, worst in 19 years. *The Hindu*.

Cui, H., Mittal, V. O., & Data, M. (2006). Comparative Experiments on Sentiment Classification for Online Product Reviews. *American Association for Artificial Intelligence*.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Datavolume 2*, *5*.

Hearst, M. (2003). What Is Text Mining? *SIMS*.

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! *The {AAAI} Press*.

Kuhn, G. (2017). The What, How, and Why of Voice of Customer (VoC) Market Research. *Driveresearch*.

Liu, B. (2010a). Sentiment analysis: A multi-faceted problem. *IEEE Intelligent Systems*.

Liu, B. (2010b). *Sentiment analysis and subjectivity*.

Luca, M. (2011). Reviews, Reputation, and Revenue: The Case of Yelp.Com. *SSRN Electronic Journal*.

Mehra, J. (2020). Hyundai Venue sales in India cross 97,000 units in first year. *Autocar*.

Nadkarni, A. (2020). Will the 2020 Hyundai Creta outsell the Kia Seltos? *Team BHP*.

Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*.

Ranjan, N., Gupta, A., Dhumale, I., Gogawale, P., & Gramopadhye, R. (2015). A SURVEY ON TEXT ANALYTICS AND CLASSIFICATION TECHNIQUES FOR TEXT DOCUMENTS. *International Journal of Development Research*, *5*(11), 4.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using Text Mining Techniques for Extracting Information from Research Articles. *Studies in Computational Intelligence*.

Singh, A. K., Gupta, D. K., & Singh, R. M. (2017). Sentiment Analysis of Twitter User Data on Punjab Legislative Assembly Election. *International Journal of Modern Education* *and Computer Science*.

Team BHP. (n.d.). *About Us*. Team BHP.

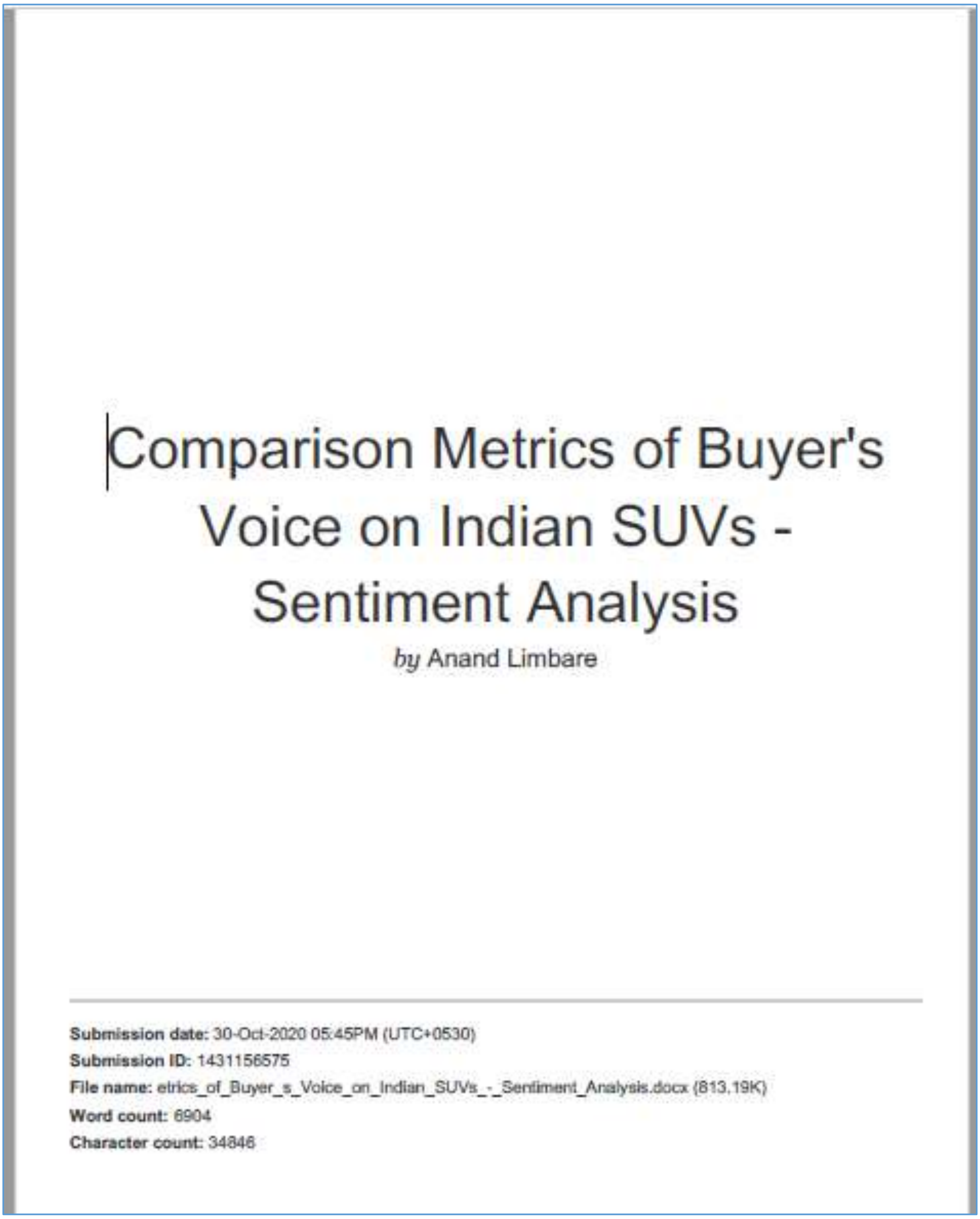
Verma, A. (2019). *Indian auto industry 2.0 – Innovation, NPD and globalisation imperatives*.

Zappa, D., Borrelli, M., Clemente, G. P., & Nino, S. (2019). Text Mining in Insurance: From Unstructured Data to Meaning. *VarianceJournal*.

Zhang, W., Xu, H., & Wan, W. (2012). Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications*. https://www.researchgate.net/publication/257404010_Weakness_Finder_Find_product_weakness_from_Chinese_reviews_by_using_aspects_based_sentiment_analysis

Appendix

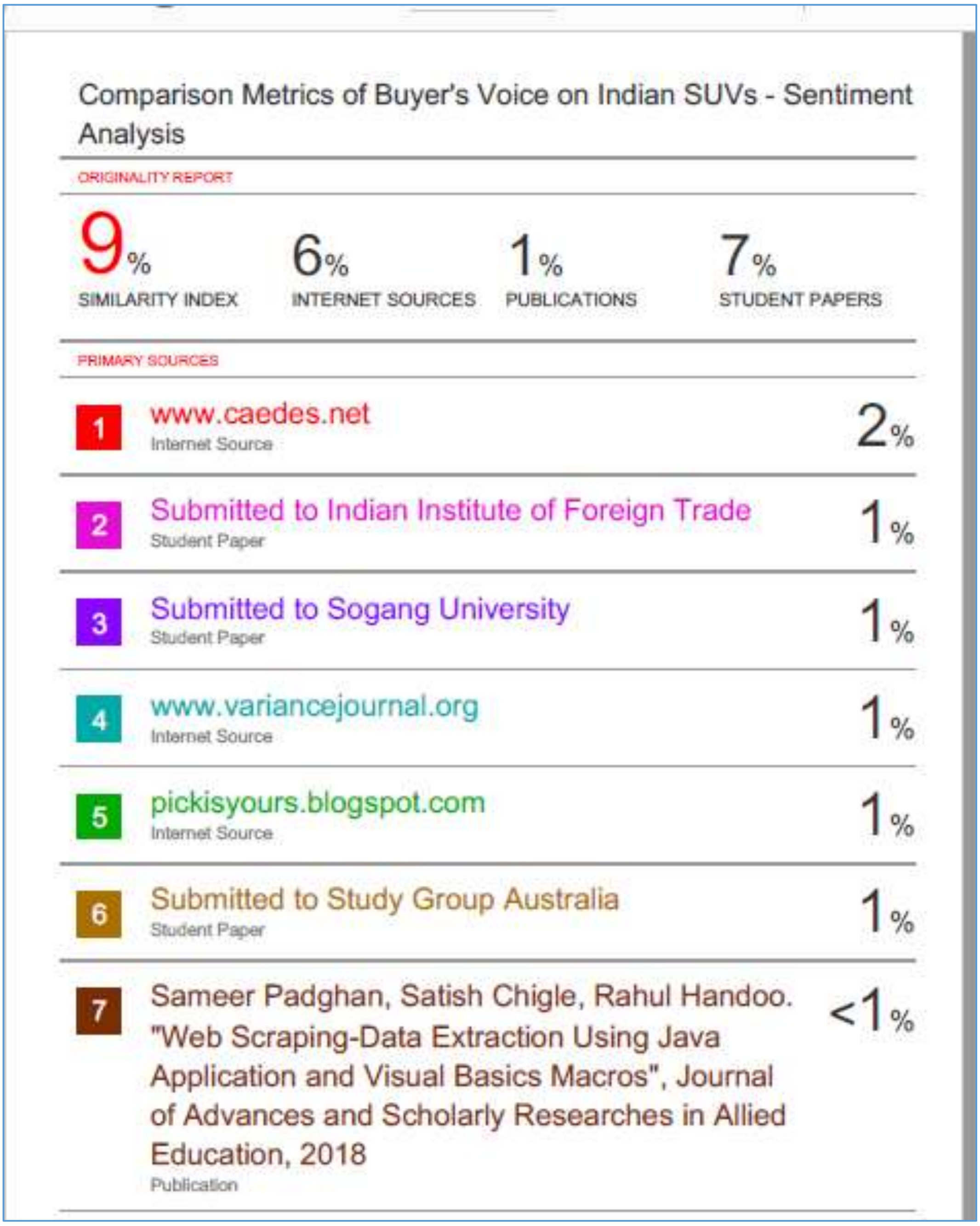
Plagiarism Report



Comparison Metrics of Buyer's Voice on Indian SUVs - Sentiment Analysis

by Anand Limbare

Submission date: 30-Oct-2020 05:45PM (UTC+0530)
Submission ID: 1431156575
File name: etrics_of_Buyer_s_Voice_on_Indian_SUVs_-_Sentiment_Analysis.docx (813.19K)
Word count: 6904
Character count: 34846



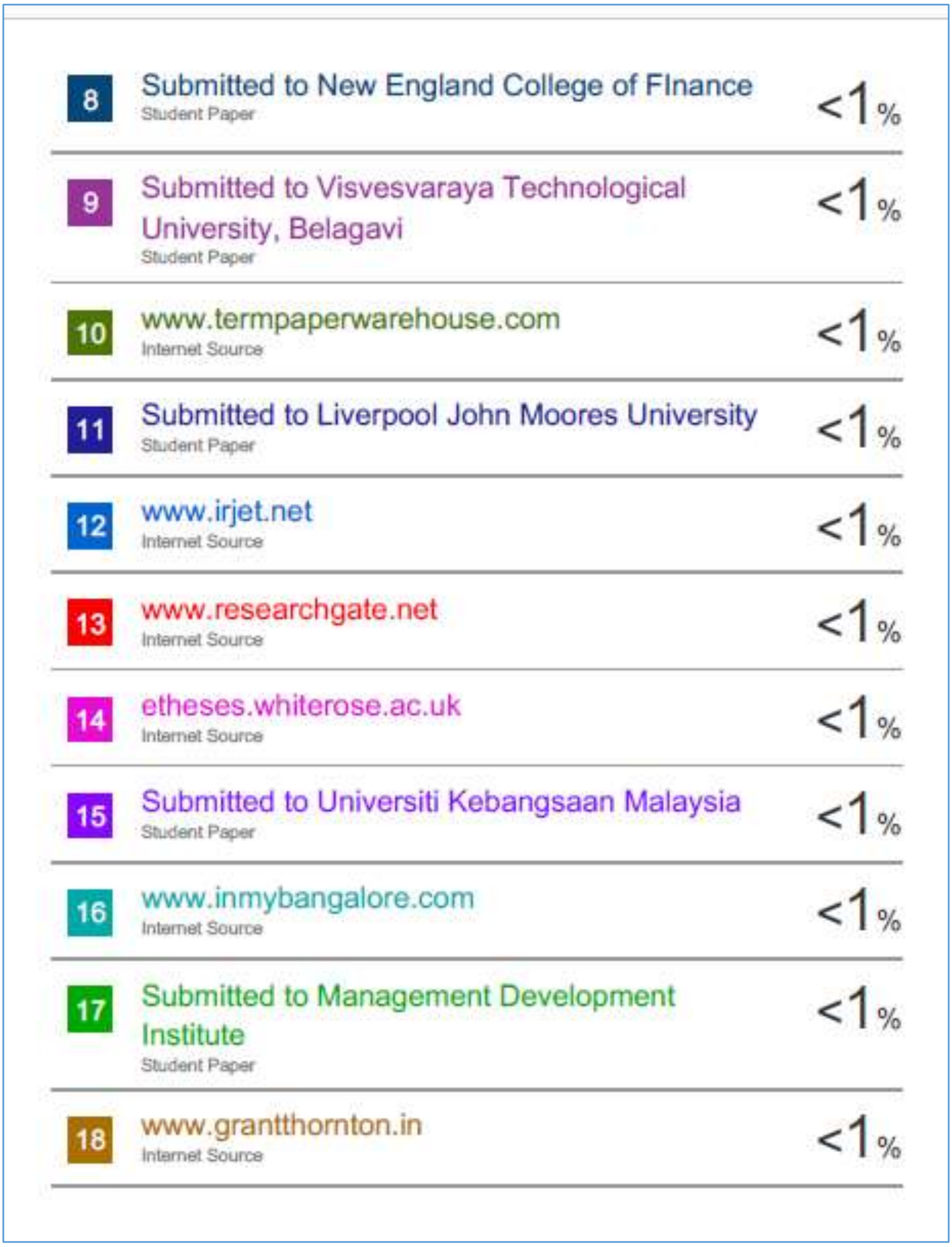
Comparison Metrics of Buyer's Voice on Indian SUVs - Sentiment Analysis

ORIGINALITY REPORT

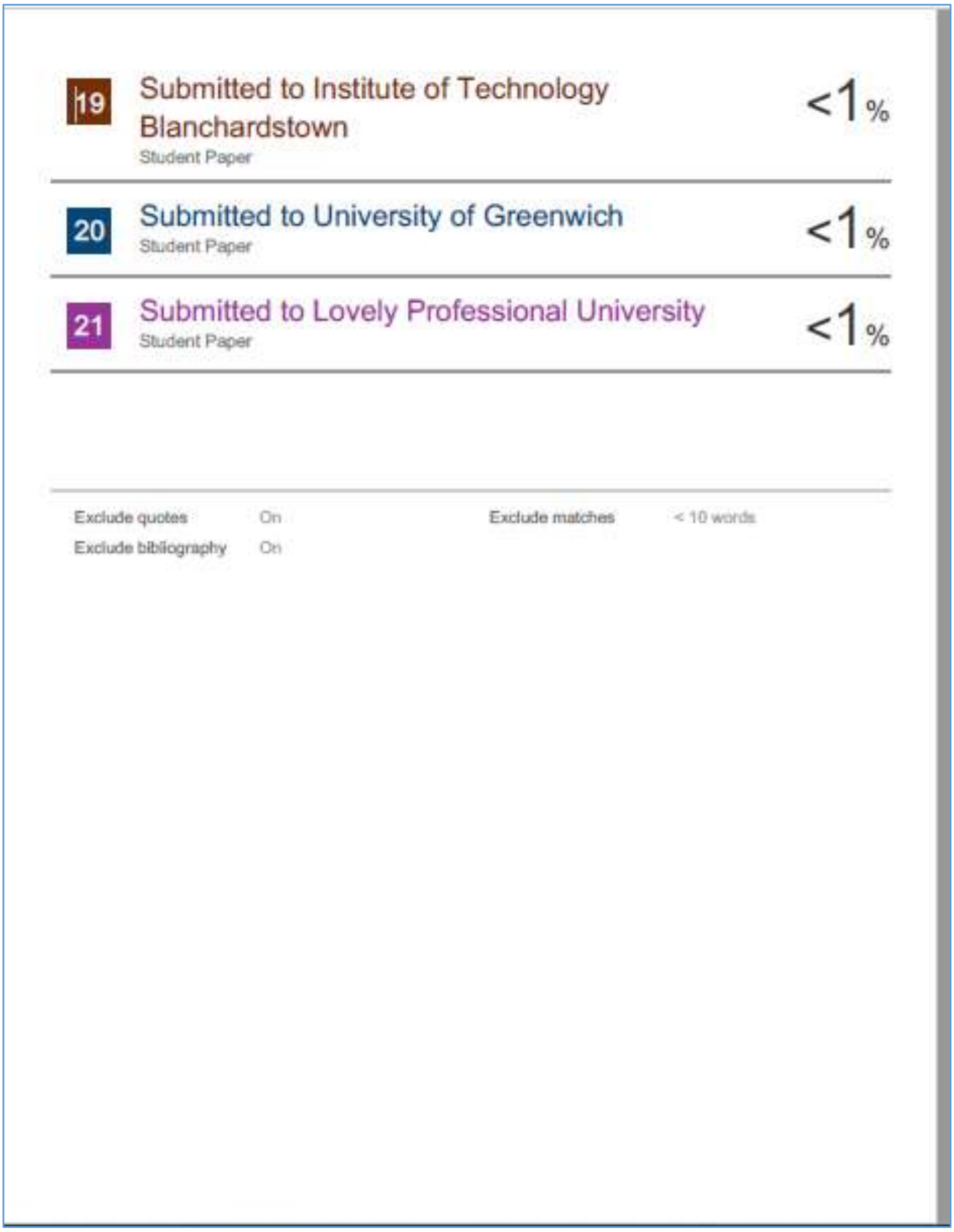
9%	6%	1%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.caedes.net Internet Source	2%
2	Submitted to Indian Institute of Foreign Trade Student Paper	1%
3	Submitted to Sogang University Student Paper	1%
4	www.variancejournal.org Internet Source	1%
5	pickisyours.blogspot.com Internet Source	1%
6	Submitted to Study Group Australia Student Paper	1%
7	Sameer Padghan, Satish Chigle, Rahul Handoo. "Web Scraping-Data Extraction Using Java Application and Visual Basics Macros", Journal of Advances and Scholarly Researches in Allied Education, 2018 Publication	<1%



8	Submitted to New England College of FInance Student Paper	<1%
9	Submitted to Visvesvaraya Technological University, Belagavi Student Paper	<1%
10	www.termpaperwarehouse.com Internet Source	<1%
11	Submitted to Liverpool John Moores University Student Paper	<1%
12	www.irjet.net Internet Source	<1%
13	www.researchgate.net Internet Source	<1%
14	etheses.whiterose.ac.uk Internet Source	<1%
15	Submitted to Universiti Kebangsaan Malaysia Student Paper	<1%
16	www.inmybangalore.com Internet Source	<1%
17	Submitted to Management Development Institute Student Paper	<1%
18	www.grantthornton.in Internet Source	<1%



19	Submitted to Institute of Technology Blanchardstown Student Paper	<1%
20	Submitted to University of Greenwich Student Paper	<1%
21	Submitted to Lovely Professional University Student Paper	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		