REVA UNIVERSITY

Bengaluru, India

A Project Report on

Identifying Voice Of Customers for Automotive Gadgets using Twitter/Facebook User Comments

Submitted in partial fulfilment for award of degree of

Master of Business Administration

In **Business Analytics**

Submitted by

Suresha K

SRN R19MBA11

Under the Guidance of

Akshay Kulkarni

Lead Data Scientist

REVA Academy for Corporate Excellence

REVA University

Rukmini Knowledge Park, Kattigenahalli,

Yelahanka, Bangalore – 560064

October, 2020

REVA
UNIVERSITY
Bengaluru, India

Candidate’s Declaration

I, **Suresha K** 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 **Identifying Voice Of Customers for Automotive Gadgets using Twitter/Facebook User Comments** under the supervision of **Akshay Kulkarni**. 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
Date: 25 Oct. 20

Name of the Student: **Suresha K**
Signature of Student

REVA
UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the Project work entitled **Identifying Voice Of Customers for Automotive Gadgets using Twitter/Facebook User Comments** carried out by **Suresha K** with **SRN R19MBA11**, is a bonafide student of REVA University, is submitting the first year project report in fulfilment for the award of **Master of Business Administration 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
Name of the Guide	Name of the Director
Guide	Director

External Viva

Names of the Examiners

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

Place: Bengaluru

Date:

REVA
UNIVERSITY
Bengaluru, India

Acknowledgement

I am highly indebted to Dr. Shinu Abhi, Director, Corporate Training for the guidance and support provides during throught the course and my project.

I would like to thank Mr. Akshay Kulkarni for the valuable guidance provided as my project guide to understand the concept and in executing this project.

It is my gratitude towards our Chief Mentor, Dr. Jay Bharateesh Simha and all other mentors for the valuable guidance and suggestions in learning various data science aspects and for the support. I am grateful to them for the valuable guidance on a number of topics related to the project.

I am thankful for my class mates for their support, suggestions and friendly advice during the project work; also I would like to thank Ameen and Saket for their help especially in Python programming.

I would like to acknowledge the support provided by Hon’ble Chancellor, Dr. P Shayma Raju, Vice Chancellor, Dr. K. Mallikharjuna Babu, and Registrar, Dr. M. Dhanamjaya.

It is sincere thanks to all members of program office of RACE who were always supportive in all requirements from the program office.

It is my sincere gratitude towards my parents, and my family for their kind co-operation. Their encouragement also helped me in completion of this project.

Place: Bengaluru
Date: 25 Oct. 20

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

This is to certify that this project report titled **Identifying Voice Of Customers for Automotive Gadgets using Twitter/Facebook User Comments** was scanned for similarity detection. Process and outcome is given below.

Software Used: **Turnitin**
Date of Report Generation: **25 Oct. 20**
Similarity Index in %: **6%**
Total word count: **9,763**
Name of the Guide: **Akshay Kulkarni**

Place: Bengaluru
Date:

Name of the Student: **Suresha K**
Signature of Student

Verified by:

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No.	Abbreviations	Long Form
1	RACE	REVA Academy for Corporate Excellence
2	VOC	Voice Of Customers
3	NLP	Natural Language Processing
4	ML	Machine Learning
5	DL	Deep Learning
6	AI	Artificial Intelligence
7	API	Application Programming Interface
8	CRISP-DM	CRoss Industry Standard Process for Data Mining
9	NLTK	Natural Language Toolkit
10	TF-IDF	Term Frequency Inverse Document Frequency
11	NLG	Natural Language Generation
12	EDA	Exploratory Data Analysis
13	BOW	Bag Of Words
14	AFINN	Årup Finn Nielsen
15	POS	Part Of Speech
16	EDA	Exploratory Data Analysis
17	SVC	Support Vector Classifier
18	NB	Naïve Bayes
19	NLTK	Natural Language Tool Kit
20	VADER	Valence Aware Dictionary and sEntiment Reasoner
21	TF-IDF	Term Frequency Inverse Document Frequency
22	CNN	Convolutional Neural Networks
23	RNN	Recurrent Neural Networks
24	Word2Vec	Word to Vector
25	GloVe	Global Vectors for Word Representation
26	BERT	Bidirectional Encoder Representations from Transformers
27	LSTM	Long Short Term Memory
28	GRU	Gated Recurring Units
29	RNN	Recurrent Neural Network

30	SWOT	Strengths Weaknesses Opportunities Threats
31	API	Application Programming Interface

List of Figures

List of Tables

Abstract

With the arrival of text analytics, Voice of Customer data become a crucial resource which provides the managers and marketing practitioners with consumer’s indirect opinion and requirements.

The use of VOC data improves the customer responsiveness and satisfaction and entually improves business performance.

It can be used for predicting service time based on voice of customer data.

I have studied customer sentiments on social media for the gadgets of automobiles.

I have used Lexicon based and Machine Learning approaches for the social media sentiment analysis.

The customers use social network sites like Twitter, Facebook, Instagram etc. so on as they offer very user friendly environment and comfort, user express their views freely. The scientists and researchers can use the social media platforms to get the data and can understand the customer behaviour and their sentiment towards the products and brands. The consumers’ comments and expressions over social networks can directly affect the brand.

The companies also adopt social networks for their product promotions and Twitter and Facebook are the most widely used social media platform by them. The consumers react, complain and appreciate them on the comments which will be valuable data for the companies. Customers also spend more time in online marketing rather the actual one. The requirement is here is to identify what customer enjoys the most of the products and revisit.

The sentiment analysis is classifying the polarity of the text or sentence to be positive, negative, or neutral. Sentiment classification also makes uses of the emotional states like "angry", "sad", and "happy".

I would like to use Text Analytics modelling approach to find the VOC from social media. The data has been extracted by web scrapping and to combine the voice of customers from social media like Facebook, Twitter and web scrapping from other platforms. The research will help the companies to get the voice of customer from multiple social media and can consider as input for business decision making on their products and commercializing the product for promotional and aftermarket voice of customers.

Here, I want to propose a dashboard of consumer sentiment for the Automotive Gadgets from Twitter and Facebook and finally an API which helps the enterprises for their market research, virtual market surveys and to identify voice of customers.

Keywords: Text Mining, Sentiment Analysis, Natural Language Processing, Voice Of Customer, Word Embedding, Facebook, Twitter

Contents

Chapter 1: Introduction

With increased competition and continually growing capital, it is increasingly important for the automotive industry to recognize well-defined market opportunities before committing resources to developing a new product in the area of Infotainment Gadgets. Creating market-related quality products can mean the difference between large profits and large losses. The intense competition makes it increasingly difficult to achieve success. Target marketing analytics provides a means to better identify opportunities that are more likely to succeed because they are based on consumer data, not just observation and interpretation. Strictly matching demographic and social economic backgrounds in today's market is not enough to effectively market technology items for automotive infotainment. The expectation and requirement of marketplace has been quick and high and there are competitors who satisfy the need quickly too. Therefore, one-dimensional analysis of non-emotive variables is not enough even for such an approach, the modern automotive infotainment market is far too sophisticated. Numerous customer surveys and product feedback are available from Twitter nowadays since the consumers are very likely to post their experience about the product over the Internet.

It is a vital undertaking to collect this data, separate client evaluations and arrange them. Virtual marketing task takes the advantage of NLP to analyse huge amount of Tweets in order to gather the opinion about the product posted by different users.

The main objective of this project is to study the sentiment analysis methods of Twitter and Facebook data and provide voice of customers on automobile gadgets. Different sentiment-analysis approaches were defined, including supervised, unsupervised and hybrid approaches used for Twitter data. Lastly, the latter's discussions and comparisons are emphasized.

My area of study here is like Decision tree classifier, SVC algorithm, Random Forest Classifier, Gradient Boosting Classifier and Gaussian Naïve Bayes for the sentiment analysis and techniques like Word Embedding for NLP.

In this field, multiple researchers have done different work. events such as detection of earthquakes using social sensors , summarization of events, interpretation of public feelings on Twitter, etc. (Pak & Paroubek, 2010). As time goes on, these are all the advances in research. Analysis of feelings has therefore become a popular field of research work.

Here, I have used major Text mining architecture and NLP framework to solve the problem.

The spell correction and word embedding techniques have been made the study unique and the final dashboard and API has made this study proposal and real market requirement.

In Text mining NLP approach negation is an important aspect as they affect polarity of other words. The words like no, not, shouldn’t etc.

Negations affect the polarities of words. Hence it is important to treat negations. They include words like no, should not, not etc. NLP analyses large amount of text data and describes how to program computers to process.

It is a subfield of linguistics, computer science, and artificial intelligence.

Python has many libraries for natural language processing like NLTK, TextBlob, Gensim, spaCy and so on. To process the natural language text and to extract the useful information from them, the sentence needs to be converted into a set of numbers or vectors as a requirement for machine learning and deep learning. This is called Word Embedding.

Chapter 2: Literature Review

"NLP is one of the significant topics that concerns about the interrelation among the huge amount of unstructured text on social media. By seeking information regarding how the computer systems are examining and getting information from the languages of human beings to create applications of high quality is the main goal of NLP." (Salloum et al., 2017)

The target for the Virtual Market Survey while conducting sentiment analysis on Tweets is essentially to correctly identify the Tweets in different classes of sentiment. Different approaches have developed in this field of research which recommend methods to train a model and then test it to verify its effectiveness. (Azizah et al., 2017)

Digital consumer survey refers to the study of emotions that has been treated as a natural language processing task at many granularity scales (Hu et al., 2017) Sentiment Classification can be achieved by classifying a function as positive as well as negative as positive as well. (Sinha, 2018) (Singh & Kaur, 2015) (Sinha, 2018) "A corpus-based approach is based on assigning each word to the emotional affinity and then identifying each of them from the huge corpus to the probabilistic score." (Shukri et al., 2015) Approach based on machine learning uses the technique of classifying text into categories.

"Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses and they do not provide with the correct targets at all and therefore rely on clustering." (Win et al., 2017)

"Supervised learning: It is based on labelled dataset and the labels are provided to the model during the process. These labelled datasets are trained to get meaningful outputs when encountered during decision making." (Win et al., 2017)

"A bag-of-words is a representation of text that explains the occurrence of words within a document. The occurrence of words is represented in a numerical feature. It is a way of extracting features from the text for use in modelling, such as with machine learning algorithms." (Pawar et al., 2015) "The confusion matrix gives the better understanding of how correctly the Tweets are classified into pre-fined classes." (Pawar et al., 2015)

In this project, my focus area is to understand and to identify the voice of customers by extracting Facebook and Twitter user comments; by sing text pre-processing and NLP techniques to identify VOC. Here, ultimate aim is to develop state of the art dashboard and API as future scope. This front end will serve as master source for the enterprises to identify the sentiments of customers and to take business decisions.

Chapter 3: Problem Statement

Identify voice of customers for automotive gadgets quickly using user comments in Twitter and Facebook and give the sentiments of customers by which enterprises take business decision making on their products.

Approx. 64% of the respondents used a smartphone application to assist with their travel, Navigation and Real-time Traffic Information Systems, Safety, Bluetooth, In-Vehicle Technology (Shukri et al., 2015). 100 million users generating over 500 million tweets every day.

"According to Edelman (2007), customers are currently switching to usage of social networks and are spending much more time with online marketing than with Contemporary Management Research 76 any other marketing channel." (Ramsaran-Fowdar, 2013) "Facebook allows companies to connect with many more people and much more often than the companies would be able to approach through phone calls, emails, or meetings" (Ramsaran-Fowdar, 2013)

It is imperative for automotive companies too to use and implement latest technologies to get the sentiments and mining the user opinions from social media. The traditional market research, surveys and getting voice of customers antiquated and need for the use of AI / ML techniques is more essential today.

In this project, I have used the NLP techniques of Text Mining to solve the problem stated regarding analysing the sentiments and to get voice of customers.

Facebook and Twitter – both have their own advantages and limitations. Twitter can be considered for quicker feedback and the hashtags become more trending where there are more sentiments. However it has the limitation of text to 280 characters. Another point needs to be considered here is the re-Tweets. Most trending hashtags see more re-Tweets which can be taken as highly emotional, however needs to be careful of fake accounts and it is difficult for machine to understand sarcasm also.

Facebook being seen with more engagement from the users can be seen with more detailed feedback. The product companies can see the user comments in their Facebook pages when they launch their new products, when they release products with new features and so on.

Customers express their happiness, concerns and also their expectations. This makes an advantage also to the companies they can focus on research and development area for the improvised product features in the next release. The comments can be seen with other user sentiments also with their emojis, replies with their approval or an additional experience which should be significant as customer feedback.

The user comments are unstructured and we can expect a lot of pre-processing required for the data to get a meaningful data for the analysis. We can see spelling mistakes, handles like @ symbol, hasthtags, emojis, repeated words and so on. Hence text pre-processing is a significant step in Text Mining. I have used NLP for my project here.

The problem which I am going to address in this project is to “Identify voice of customers for automotive gadgets using user comments in Twitter and Facebook and to know the sentiments of customers by which enterprises can do business decision making for market research of their product”

Chapter 4: Objectives of the Study

As mentioned earlier, Twitter and Facebook are the most used social media platforms today across the globe. The end customers of any product today - automotive gadgets in this project, is a user of anyone of these or both. They express their sentiments – gladness, sad, concerns or any issues with the product in their timeline or in the Facebook page or in Twitter handles of companies of respective products.

The objective wants to see different NLP techniques and approaches to do the pre-processing of text and identify the best algorithm which gives highest accuracy.

The study has been done with the existing NLP practises through available research papers in the area and has been taken as reference for my study. As converting text to features is an important step in Text mining, I have used different techniques of feature extraction like TF-IDF, TextBlob and also Word2Vec.

The final objective of the study is to develop a front end API as a front end web application where clients can use the API for their market research, market surveys and to get voice of customers virtually and quickly.

Chapter 5: Project Methodology

Framework: “The CRISP-DM framework has been used here for the project. It is a process model which explains approaches in data mining. It is most widely used model in analytics. It was conceived in 1996. IBM Corporation released a new methodology called Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM). It refines and extends CRISP-DM. It consists of following 6 steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. The project will be explained in these 6 steps in the following pages.” (Wikipedia, 2020)

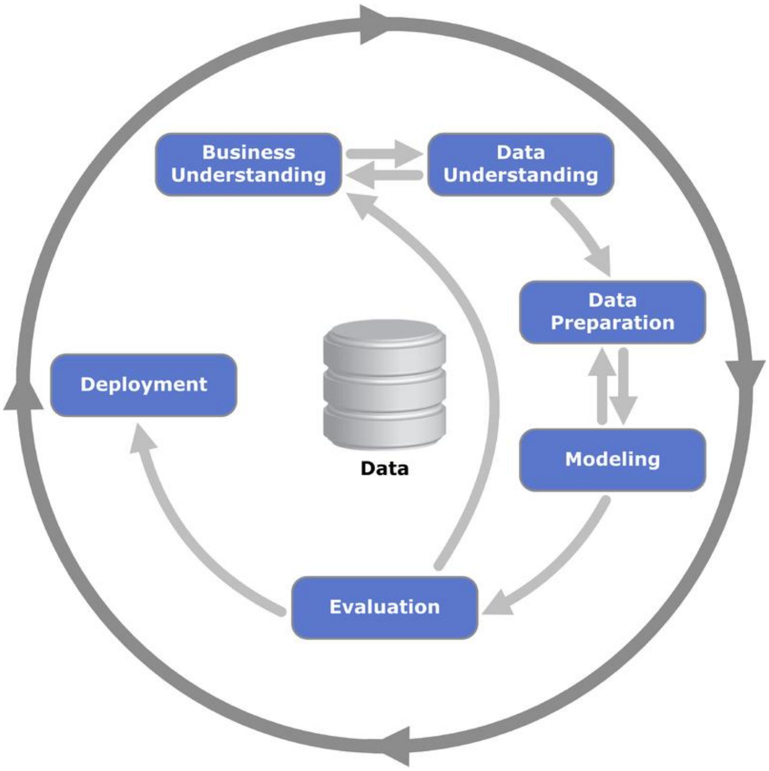


Figure 5.1 CRISP-DM Framework (Wikipedia, 2020)

Business Understanding:
“It emphasizes on understanding the objectives of a project and its requirements from a business perspective. It converts this knowledge into a data mining problem definition and initial plan.”(Wikipedia, 2020)
Data Understanding:
“It starts with collection of data and conducts activities to get familiar with the data to identify data quality problems and discovers insights into the data or detects hidden information.” (Wikipedia, 2020)
Data Preparation:
“This phase covers the activities to construct the dataset from initial raw data. This will be the final dataset for modeling.” (Wikipedia, 2020)
Modeling:

“Various ML / DL modeling methods are selected and applied, there can be a loop back to Data Preparation since some modeling techniques require specific requirements form of data.” (Wikipedia, 2020)
Evaluation:
“In this phase, we will assess the degree to which the model meets our business objectives.” (Wikipedia, 2020)
Deployment:
“In this phase of deployment, we will take our evaluation results and determine a strategy for their deployment. It can be seen the loop back in Data Understanding, Modeling and Evaluation phases.” (Wikipedia, 2020)

Text analytics is the process of converting unstructured text to structured form and applying statistical/AI methods to discover hidden patterns, predict the classes/labels as required. In addition summarization to understand the text faster is also a part of text analytics. The process of Text analytics has four stages. The unstructured data is pre-processed and converted into standard format for analysis. It is the first stage. The different models are developed for various requirements as clustering, classification, Information extraction (IE), Topic Detection and summarization. It is the second stage. In the third stage, the discovered knowledge is presented as per the requirements. In the fourth stage, the developed models will be used for scoring unseen data. The fourth step requires first and second step, similar to third step of presentation.

NLP Approach: NLP is my focus area in this project.

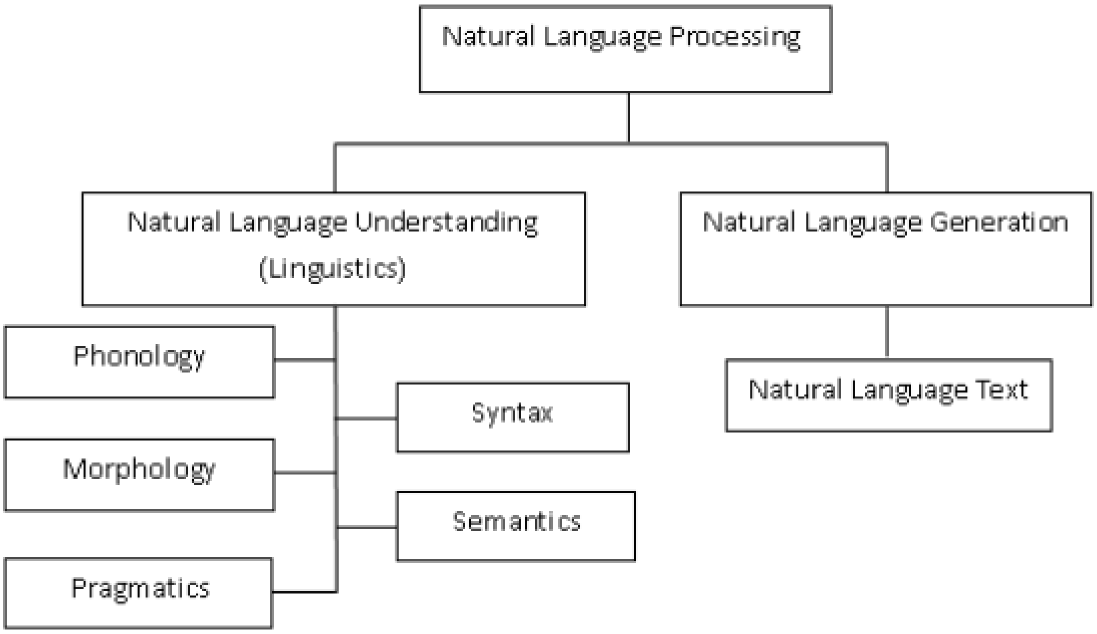

Natural Language Processing
Natural Language Understanding (Linguistics)
Natural Language Generation
Phonology
Syntax
Natural Language Text
Morphology
Semantics
Pragmatics

Figure 5.2 Broad Classification of NLP (Khurana et al., 2018)

“Linguistics is the science of language. It includes Phonology that refers to sound, Morphology refers to word formation, syntax sentence structure, Semantic Syntax and Pragmatics refers to understanding.” (Khurana et al., 2018)

“NLG is the process pf producing phrases, sentences and paragraphs. They are meaningful from an internal representation. It has four phases: identifying the goals, planning on how goals maybe achieved by evaluating the situation and available communicative sources and realizing the plans as a text. It is opposite of NLU.” (Khurana et al., 2018)

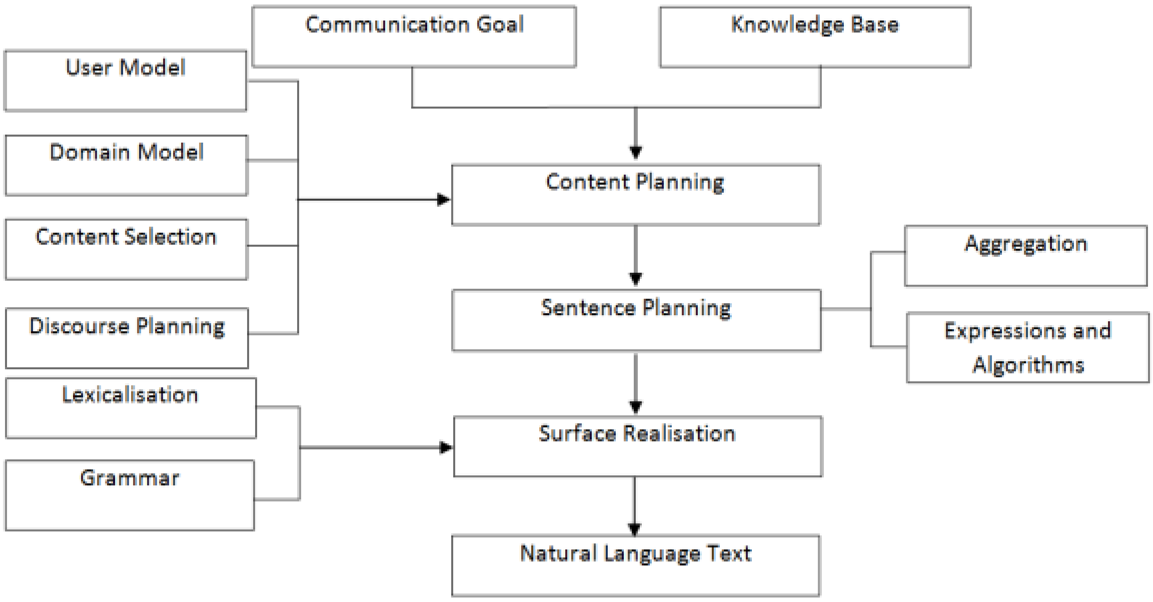


Figure 5.3 Phases of NLP Architecture (Khurana et al., 2018)

“Sentiment analyser (Yi et al., 2003) extracts sentiments about given topic. It consists of a topic feature term extraction, sentiment extraction and association by relationship analysis. It uses two linguistic resources for the analysis – the sentiment lexicon and the sentiment pattern database. It analyses the documents for positive and negative words and try to give ratings on the scale -5 to +5.” (Khurana et al., 2018)

“Parts of speech taggers for the languages are being done on making parts of speech taggers for other languages.” (Khurana et al., 2018)

“Chunking works by labelling segments of sentences. It is with syntactic correlated keywords like Noun Phrase and Verb Phrase (NP or VP). It is also known as Shadow Parsing,

Emotion Detection (Sharma et al., 2013)t works on social media platforms on mixing of two languages with English and any other Indian Language. It is similar to sentiment analysis.” (Khurana et al., 2018)

“Sematic Role Labelling – SRL It works by giving a semantic role to a sentence in the document, Event discovery in social media feeds (Benson et al., 2011) it uses a graphical model to analyse any social media feeds and it determines whether it contains name of a person or name of venue, place, time etc.” (Khurana et al., 2018)

Chapter 6: Business Understanding

“Sentiment analysis in business is also known as opinion mining is a process of identifying and making catalogue a piece of text according to the tone conveyed by it. His text can be Tweets, comments, feedback and ratings like positive, neutral and negative ass sentiments associated with it. The requirement of any business is to implement an automated sentiment analysis process. When the data is huge, manual analysis is not an option as it is tedious process. Hence sentiment analysis in business is necessary more than a trend.” (Bilyk, 2017)

“Sentiment analysis may sometimes a major breakthrough in business as it may result in a complete change in the brand of product. In business, there are many examples of success by using sentiment analysis for improving business.” (Bilyk, 2017)

“Reputation management to build the brand: The brand monitoring and reputation management is most important use of sentiment analysis. It gives how consumers perceive company’s brand/product/service. It helps to track the perception of the brand by customers, specific details about the attitude, to find any pattern and trends and to keep a close eye on the influencers.” (Bilyk, 2017)

“An example is: KFC has successfully implemented sentiment analysis. KFC was stuck in the past for a while. The competition was moving ahead. KFC started memes like riding on the waves and pop culture iconography. This was to instil the brand proposition. This generated natural traction around the brand. It augmented by the pop culture reference. That made KFC to retain and build its brand in the competition.”(Bilyk, 2017)

“Market research and competitor analysis: Sentiment analysis brings additional perspective and insights to market research along with other tools for market research. It also helps to understand and study the competitors’ brands and their efforts in in building their brands.” (Bilyk, 2017)

“An example is: Apple uses efficiently sentiment analysis for research and competitor analysis. It uses brand value proposition, addressing various issues, introducing new features, announcing milestones and so on to address the pain points on bad design, poor privacy and low battery life.” (Bilyk, 2017)

“Product Analytics: It stems from reputation management.it goes for specific comments and remarks about the product. Algorithms from sentiment analysis can do the work and show what kind of feedback comes from which segment and from which category of audience, also what it points.” (Bilyk, 2017)

“An example is: Google uses sentiment analysis for its product improvement efficiently. The development team from Google Chrome constantly monitors its user feedback and whether it is direct or indirect sentiment, specific aspects, recommendations.” (Bilyk, 2017)

“Voice of customer analysis: The key elements of the effective business operation are accurate target audience segmentation and subsequent value proposition formulation. It is to keep a pulse of the customer in order to remain relevant and to keep the product in demand.” (Bilyk, 2017)

“An example is: VOC done by TripAdvisor. It applies aspect based sentiment analysis in order to take most out of large amount of data it generates. It allows extracting the feasible points regarding customer feedback and service.” (Bilyk, 2017)

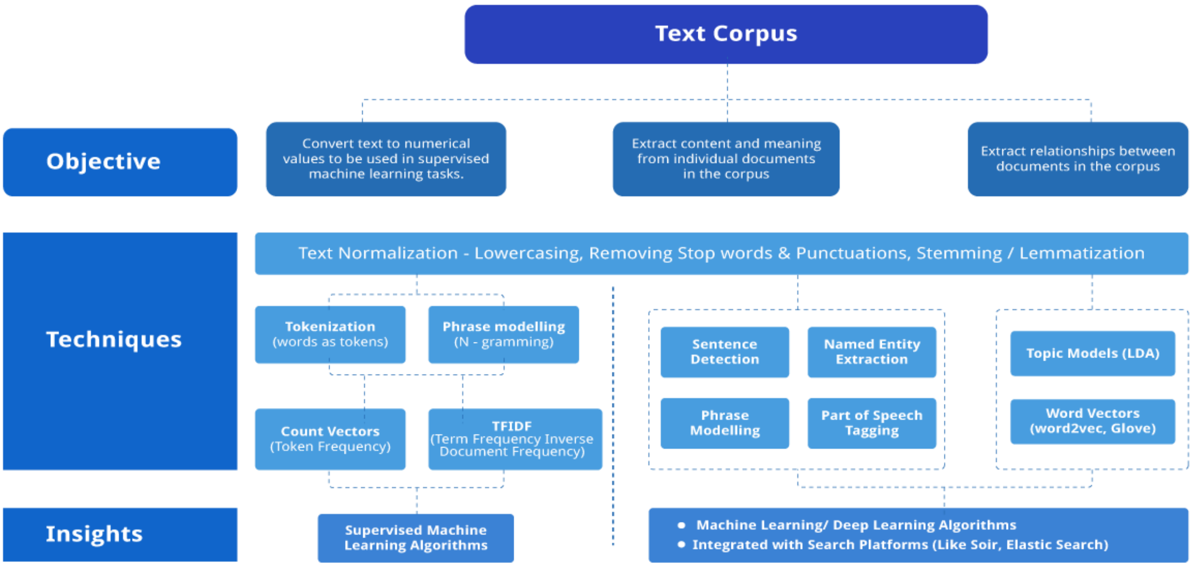
“Customer support – feedback analysis: It is one of the important elements of sentiment analysis application in real life. The different ways to implement are, insight into customer’s opinion on the product, intent analysis for process automation and workflow management and customer prioritization.” (Bilyk, 2017)

An example is: Using sentiment analysis in customer support can be seen in big tech companies. (Bilyk, 2017)

In this project, I am focussing on the business opportunity as identifying voice of customers for automobile infotainment gadgets from Twitter and Facebook user comments. The commercial benefit of this project is, the end product will be a front end API dashboard which tells the customers about the sentiment and customer feeling on their product and brands, which help them to take efficient business decisions.

Chapter 7: Data Understanding

The typical NLP architecture has been used here in this project. As shown in below flow chart image, the data input has been taken from the user comments from Facebook and Tweets from Twitter.



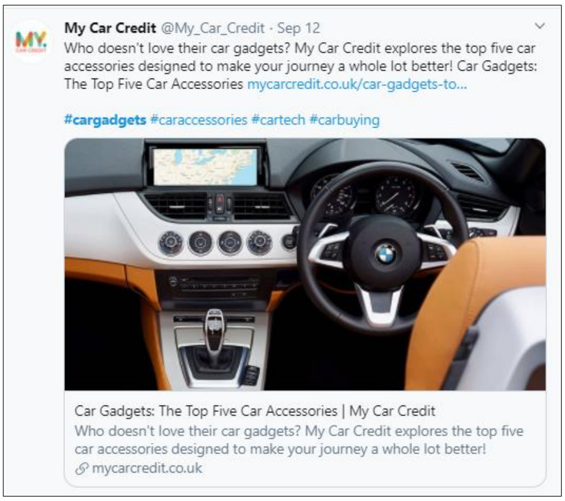
Text Corpus
Objective
Convert text to numerical values to be used in supervised machine learning tasks.
Extract content and meaning from individual documents in the corpus
Extract relationships between documents in the corpus
Techniques
Text Normalization - Lowercasing, Removing Stop words & Punctuations, Stemming / Lemmatization
Tokenization (words as tokens)
Phrase modelling (N - gramming)
Count Vectors (Token Frequency)
TFIDF (Term Frequency Inverse Document Frequency)
Sentence Detection
Named Entity Extraction
Phrase Modelling
Part of Speech Tagging
Topic Models (LDA)
Word Vectors (word2vec, Glove)
Insights
Supervised Machine Learning Algorithms
Machine Learning/ Deep Learning Algorithms
Integrated with Search Platforms (Like Soir, Elastic Search)

Figure 7.1 NLP Flowchart (Sankaran, 2017)

The data has been acquired from Twitter API and Export Comments APIs from Facebook. The user comments are normally attached with hashtags in any social network. A hashtag is a symbol; # is a type of metadata tag used on social network like Twitter, Instagram and other microblogging platforms. It is used to tag to respective topic of discussion or opinion. It is used in Facebook user comments also. This will help other users to find the related comments easily with the hashtags. The following table shows some of the hashtags from which the user comments extracted from Twitter and Facebook. The comments are tagged with the company's name also.

#CarGadgets	#CarSpeakers	#Dashboard
#Infotainment	#CarGPS	#HUD
#AndroidAuto	#ConnectedVehicles	#DashCamera
#AppleCarPlay	#CarCockpit	#GPSNavigation
#InCarEntertainment	#JBLAudio	#SelfDriving
#CarAccessories	#DashCam	#Dashcam
#CarMusic	#CarAudioSystem	#AutonomousVehicles
#CarAudio	#CarNavigation	#AutonomousDriving
#AmazonEchoAuto	#CarCamera	#SelfDrivingCars
#CarInterior	#CarMobileHolder	#CarCharger
#Gadgets	#CarStereo	#ElectricCar
#SelfDrivingCars	#CarVideo	#CarCharging

Table 7.1 Major Hashtags



My Car Credit @My_Car_Credit · Sep 12

Who doesn't love their car gadgets? My Car Credit explores the top five car accessories designed to make your journey a whole lot better! Car Gadgets: The Top Five Car Accessories mycarcredit.co.uk/car-gadgets-to...

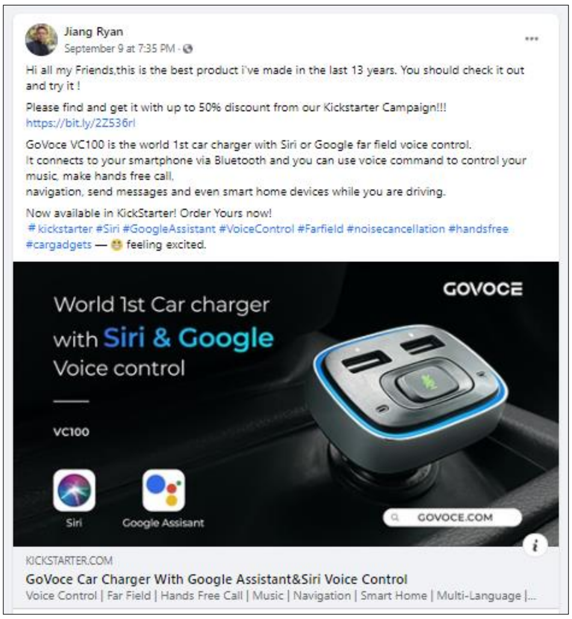
#cargadgets #caraccessories #cartech #carbuying

Car Gadgets: The Top Five Car Accessories | My Car Credit

Who doesn't love their car gadgets? My Car Credit explores the top five car accessories designed to make your journey a whole lot better!

mycarcredit.co.uk

Figure 7.2 Example of a Tweet



Jiang Ryan

September 9 at 7:35 PM ·

Hi all my Friends,this is the best product i've made in the last 13 years. You should check it out and try it !

Please find and get it with up to 50% discount from our Kickstarter Campaign!!!
https://bit.ly/2Z536rl

GoVoce VC100 is the world 1st car charger with Siri or Google far field voice control.
It connects to your smartphone via Bluetooth and you can use voice command to control your music, make hands free call,
navigation, send messages and even smart home devices while you are driving.

Now available in KickStarter! Order Yours now!
# kickstarter #Siri #GoogleAssistant #VoiceControl #Farfield #noisecancellation #handsfree #cargadgets — 😆 feeling excited.

KICKSTARTER.COM

GoVoce Car Charger With Google Assistant&Siri Voice Control

Voice Control | Far Field | Hands Free Call | Music | Navigation | Smart Home | Multi-Language |...

Figure 7.3 Example of a Facebook comment

The above figures are examples of a Tweet and comment from Facebook on automotive gadgets.

These user comments are unstructured text data as messages in social networks where users express their sentiments which are crucial for the companies as feedbacks or voice of customers who used their products. The extracted user comments have been collected as user comments column in MS Excel file from both Twitter and Facebook.

The below figure shows the data pipeline for the project:

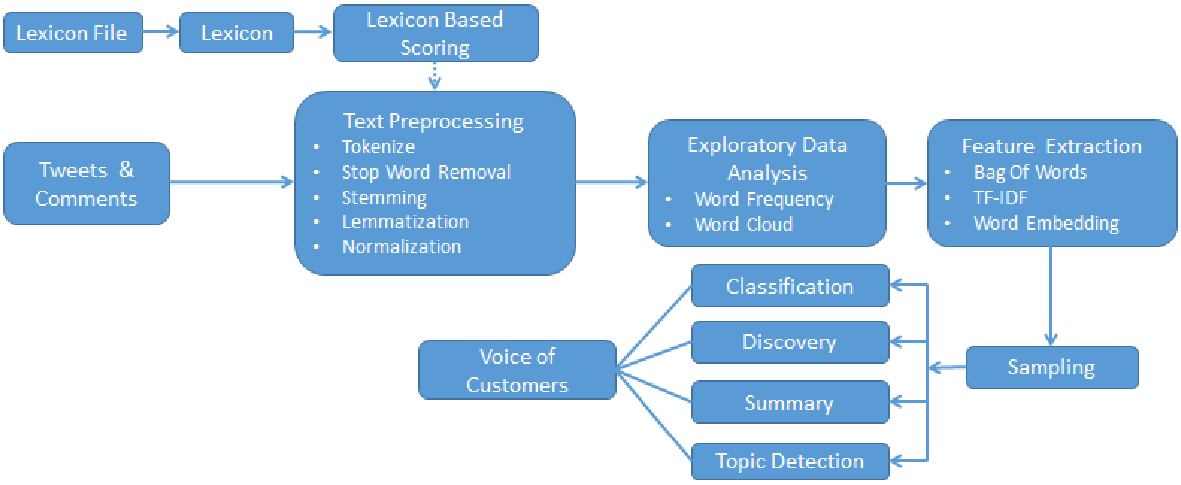
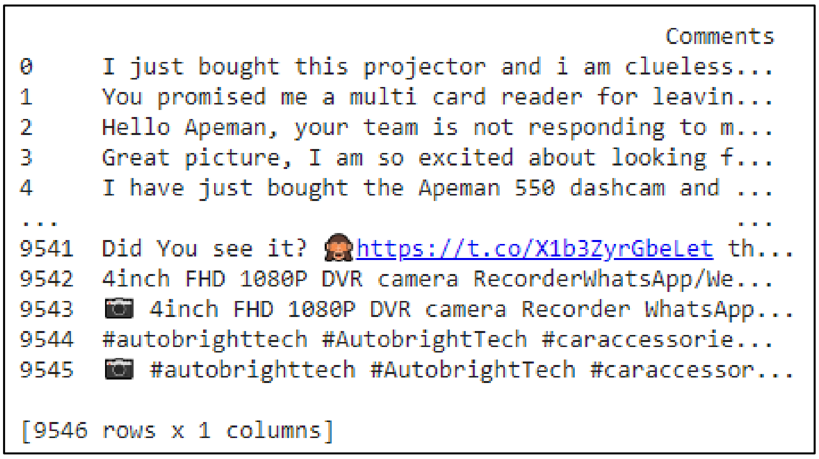


Figure 7.4 Data Pipeline

The first step would be data extracted intended for the purpose. Here, we have extracted the Tweets and Comments. In Text preprocessing, we will do various processing activities for the text. We have also used the lexicon method of scoring here. The EDA explains about the data, in this case it describes about frequency of words, repeated words and so on. After EDA, we have an important step called Feature Extraction. We need to transform the unstructured text data into the machine readable format, numbers. We have used Bag of Words, TF-IDF and Word Embedding. We then go for Sampling and then for various classifiers, discovery, Summary and Topic Detection which give the result of Voice of Customers as Positive, Neutral and Negative on the product. The detailed steps would be explained in the Data Preparation chapter.



```
                                               Comments
0      I just bought this projector and i am clueless...
1      You promised me a multi card reader for leavin...
2      Hello Apeman, your team is not responding to m...
3      Great picture, I am so excited about looking f...
4      I have just bought the Apeman 550 dashcam and ...
...                                                  ...
9541   Did You see it? 🙉https://t.co/X1b3ZyrGbeLet th...
9542   4inch FHD 1080P DVR camera RecorderWhatsApp/We...
9543   📷 4inch FHD 1080P DVR camera Recorder WhatsApp...
9544   #autobrighttech #AutobrightTech #caraccessorie...
9545   📷 #autobrighttech #AutobrightTech #caraccessor...

[9546 rows x 1 columns]
```

Figure 7.5 Data Extract

The above figure shows first 5 rows of raw data extracted by the Twitter and Export Comments APIs. We can see the Tweets and user comments for various automotive gadgets where users have expressed their views.

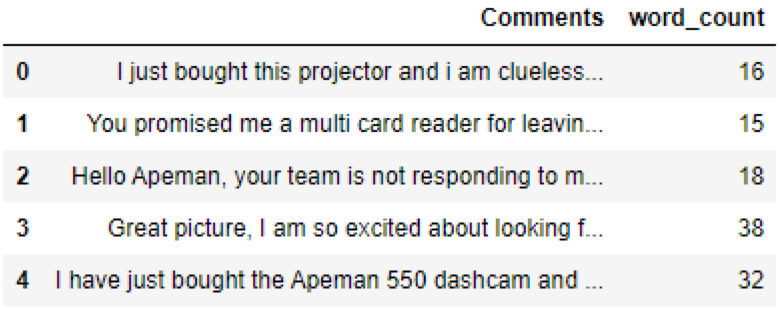
The total rows available in the dataset are 9,546 with one column for Comments. It can be seen that we have many characters in the text data which are not required for our analysis like stop words, handles like @ # symbols, emojis and so on.

Chapter 8: Data Preparation

After data collection/extract, Data Processing is an important step to make the data ready for any modeling process. In this project, the following steps involved for text preprocessing:

Basic feature extraction:

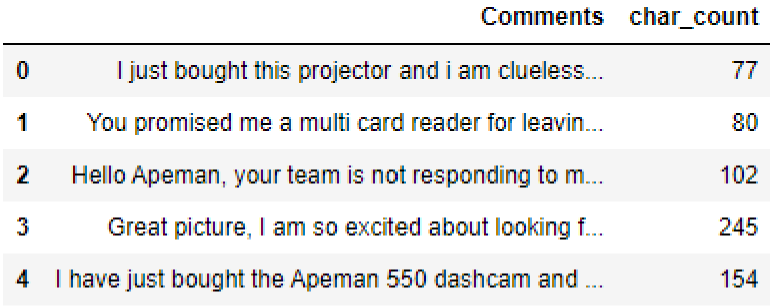
Number of words: We will see number of words in each comment extracted.



	Comments	word_count
0	I just bought this projector and i am clueless...	16
1	You promised me a multi card reader for leavin...	15
2	Hello Apeman, your team is not responding to m...	18
3	Great picture, I am so excited about looking f...	38
4	I have just bought the Apeman 550 dashcam and ...	32

Figure 8.1 Number of words

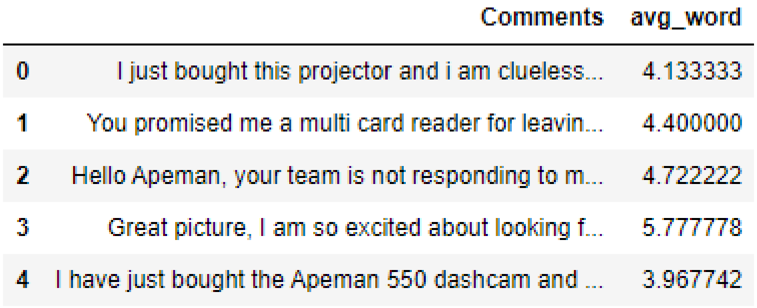
Number of characters: This tells the number of characters in each comment. We will calculate the length of each comment. It also includes the spaces which we can remove if not required.



	Comments	char_count
0	I just bought this projector and i am clueless...	77
1	You promised me a multi card reader for leavin...	80
2	Hello Apeman, your team is not responding to m...	102
3	Great picture, I am so excited about looking f...	245
4	I have just bought the Apeman 550 dashcam and ...	154

Figure 8.2 Number of characters

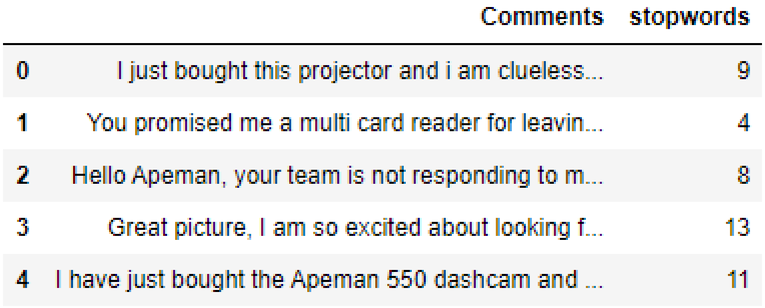
Average word length: This will calculate average word length of every comment. “We will take the sum of the length of all the words and divides it by total length of comment.”



	Comments	avg_word
0	I just bought this projector and i am clueless...	4.133333
1	You promised me a multi card reader for leavin...	4.400000
2	Hello Apeman, your team is not responding to m...	4.722222
3	Great picture, I am so excited about looking f...	5.777778
4	I have just bought the Apeman 550 dashcam and ...	3.967742

Figure 8.3 Average word length

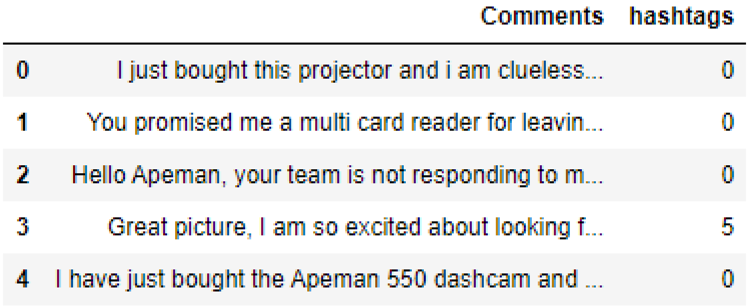
Number of stop words: "They are the most common words in a language." Some examples of stop words are, he, have etc. they are already captured in corpus. They do not add much meaning to a sentence. "We have imported stop words from NLTK, it is a basic NLP library."



	Comments	stopwords
0	I just bought this projector and i am clueless...	9
1	You promised me a multi card reader for leavin...	4
2	Hello Apeman, your team is not responding to m...	8
3	Great picture, I am so excited about looking f...	13
4	I have just bought the Apeman 550 dashcam and ...	11

Figure 8.4 Number of stop words

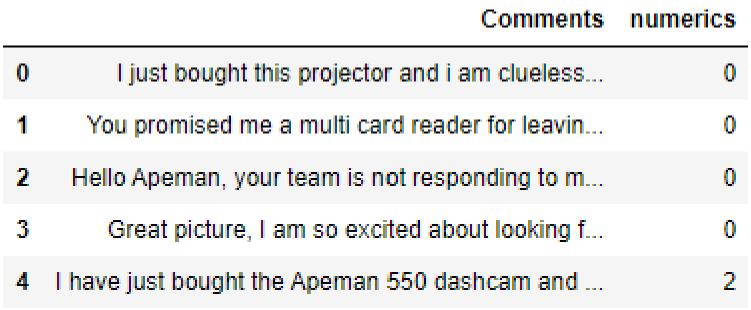
Number of special characters: The special characters like hashtags are also an interesting feature we can extract. "This will helps to additional information from text data."



	Comments	hashtags
0	I just bought this projector and i am clueless...	0
1	You promised me a multi card reader for leavin...	0
2	Hello Apeman, your team is not responding to m...	0
3	Great picture, I am so excited about looking f...	5
4	I have just bought the Apeman 550 dashcam and ...	0

Figure 8.5 Number of special characters

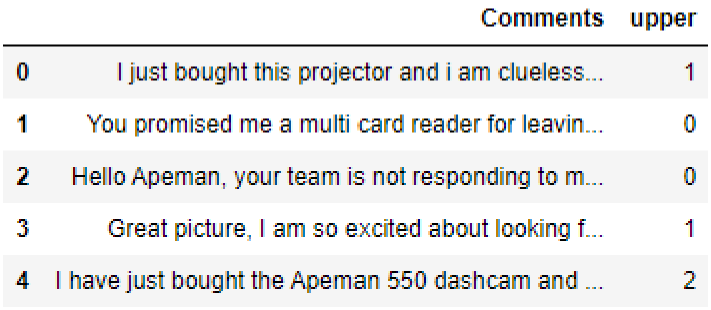
Number of numeric: This is to calculate number of numeric just like number of words. It is one of the similar exercises.



	Comments	numerics
0	I just bought this projector and i am clueless...	0
1	You promised me a multi card reader for leavin...	0
2	Hello Apeman, your team is not responding to m...	0
3	Great picture, I am so excited about looking f...	0
4	I have just bought the Apeman 550 dashcam and ...	2

Figure 8.6 Number of numeric

Number of uppercase words: This is to calculate number of uppercase words.



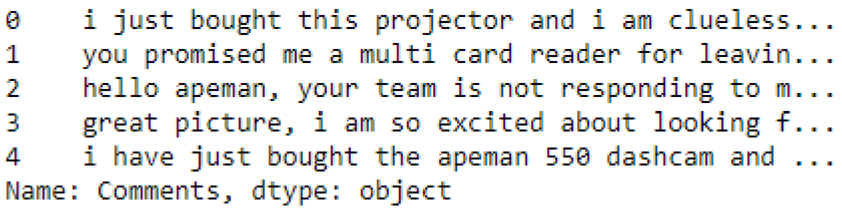
	Comments	upper
0	I just bought this projector and i am clueless...	1
1	You promised me a multi card reader for leavin...	0
2	Hello Apeman, your team is not responding to m...	0
3	Great picture, I am so excited about looking f...	1
4	I have just bought the Apeman 550 dashcam and ...	2

Figure 8.7 Number of uppercase words

Text processing:

Before moving to text and feature extraction, we need to clean the to get better features. The basic text pre-processing steps include the following:

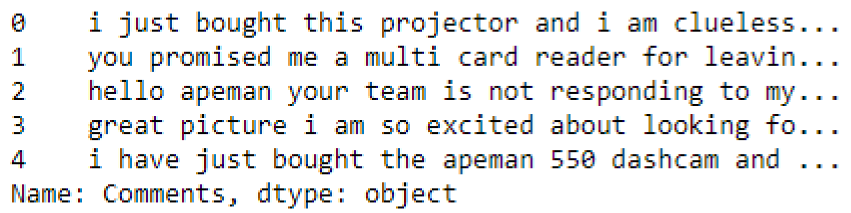
Lower case: This is used to make the sentences in lower case. “This will help to avoid multiple copies of the same words in the data.”



```
0    i just bought this projector and i am clueless...
1    you promised me a multi card reader for leavin...
2    hello apeman, your team is not responding to m...
3    great picture, i am so excited about looking f...
4    i have just bought the apeman 550 dashcam and ...
Name: Comments, dtype: object
```

Figure 8.8 Lower case

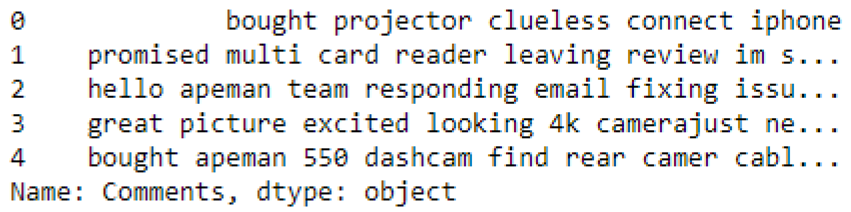
Removing Punctuations: “This step removes punctuations, as it won’t add any meaning to the text data. It will also reduce the size of the data.”



```
0    i just bought this projector and i am clueless...
1    you promised me a multi card reader for leavin...
2    hello apeman your team is not responding to my...
3    great picture i am so excited about looking fo...
4    i have just bought the apeman 550 dashcam and ...
Name: Comments, dtype: object
```

Figure 8.9 Removing punctuations

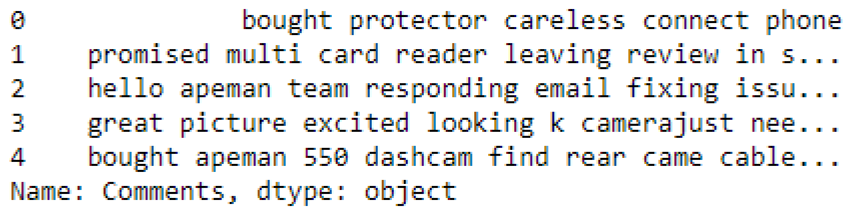
Stop words removal: As these are "commonly occurring words", they should be removed from the data. We can use predefined Python libraries to remove them. Some examples of stop words are I, me, we, our, myself etc.



```
0              bought projector clueless connect iphone
1    promised multi card reader leaving review im s...
2    hello apeman team responding email fixing issu...
3    great picture excited looking 4k camerajust ne...
4    bought apeman 550 dashcam find rear camer cabl...
Name: Comments, dtype: object
```

Figure 8.10 Removal of stop words

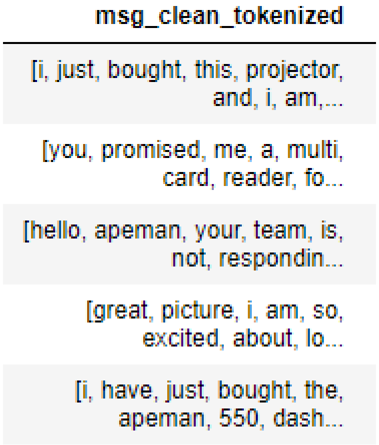
Spelling correction: As in the social network sites, we can see many spelling mistakes which we need to correct before going for modeling. We have used TextBlob library of Python here to correct the spelling.



```
0              bought protector careless connect phone
1    promised multi card reader leaving review in s...
2    hello apeman team responding email fixing issu...
3    great picture excited looking k camerajust nee...
4    bought apeman 550 dashcam find rear came cable...
Name: Comments, dtype: object
```

Figure 8.11 Spelling Correction

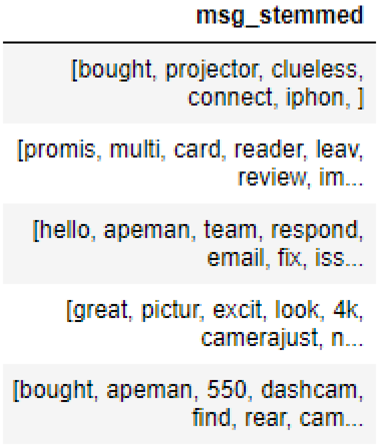
Tokenization: Tokens are building blocks of NLP. It is breaking the raw text into small words, sentences called tokens. These will help in understanding the context or developing the model and interpreting the meaning of the text. It is by analyzing the sequence of the words.



msg_clean_tokenized
[i, just, bought, this, projector, and, i, am,...
[you, promised, me, a, multi, card, reader, fo...
[hello, apeman, your, team, is, not, respondin...
[great, picture, i, am, so, excited, about, lo...
[i, have, just, bought, the, apeman, 550, dash...

Figure 8.12 Tokenization

Stemming: “It is the process of reducing a word to its word stem.” “It affixes to suffixes and prefixes or to the roots of words.” It is a normalization technique for words. For example, the stem of word “studying” is study and for the word “studies” is studi.



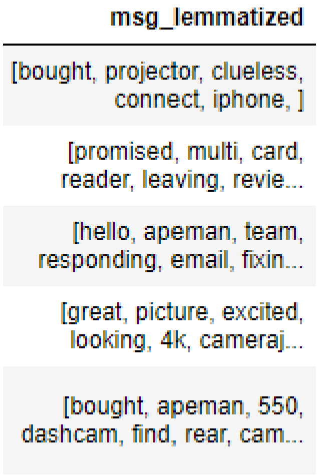
msg_stemmed
[bought, projector, clueless, connect, iphon,]
[promis, multi, card, reader, leav, review, im...
[hello, apeman, team, respond, email, fix, iss...
[great, pictur, excit, look, 4k, camerajust, n...
[bought, apeman, 550, dashcam, find, rear, cam...

Figure 8.13 Stemming

Lemmatization: It considers vocabulary and morphological analysis of words. The algorithm needs dictionaries to link the form to its Lemma. Lemma is a canonical form of a word. For example, the Lemma for the words “studying” and “studies” is study.

The difference between stemming and lemmatization is, Stemming looks at the form of the word and lemmatization looks at the meaning of the word.

There are various libraries for Stemming and Lemmatization in NLTK, TextBlob Python packages.

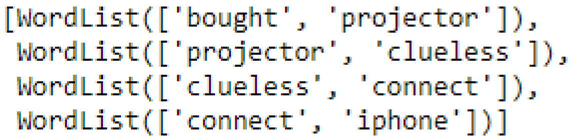


msg_lemmatized
[bought, projector, clueless, connect, iphone,]
[promised, multi, card, reader, leaving, revie...
[hello, apeman, team, responding, email, fixin...
[great, picture, excited, looking, 4k, cameraj...
[bought, apeman, 550, dashcam, find, rear, cam...

Figure 8.14 Lemmatization

Advance text processing:
We will use NLP techniques now to extract features further.

N-grams: “They are the combination of multiple words and used together.” “Unigram is when we have N=1, bigram is when we have N=2, trigram is when we have N=3 and so on.” “They capture the language structure. It captures what letter or word is likely to follow the given one.” We can work with more contexts when we have more Ns. The optimal length depends on the application.
We have used TextBlob library for this with N=2:

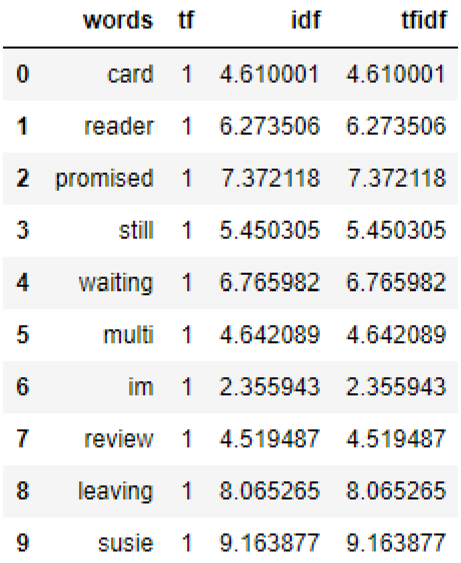


```
[WordList(['bought', 'projector']),
 WordList(['projector', 'clueless']),
 WordList(['clueless', 'connect']),
 WordList(['connect', 'iphone'])]
```

Figure 8.15 N-grams

Term Frequency: “It is measure of how frequently a term occurs in a document.
The formula is (Number of times term t appears in a document) / (Total number of terms in the document).”

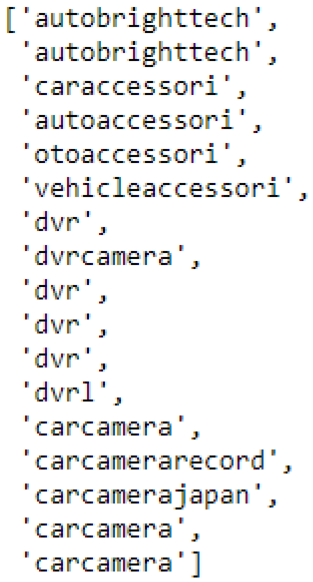
Inverse Document Frequency: “It is a measure of how important a term is.
The formula is log_e(Total number of documents / Number of documents with term t in it).”



	words	tf	idf	tfidf
0	card	1	4.610001	4.610001
1	reader	1	6.273506	6.273506
2	promised	1	7.372118	7.372118
3	still	1	5.450305	5.450305
4	waiting	1	6.765982	6.765982
5	multi	1	4.642089	4.642089
6	im	1	2.355943	2.355943
7	review	1	4.519487	4.519487
8	leaving	1	8.065265	8.065265
9	susie	1	9.163877	9.163877

Figure 8.16 TF IDF

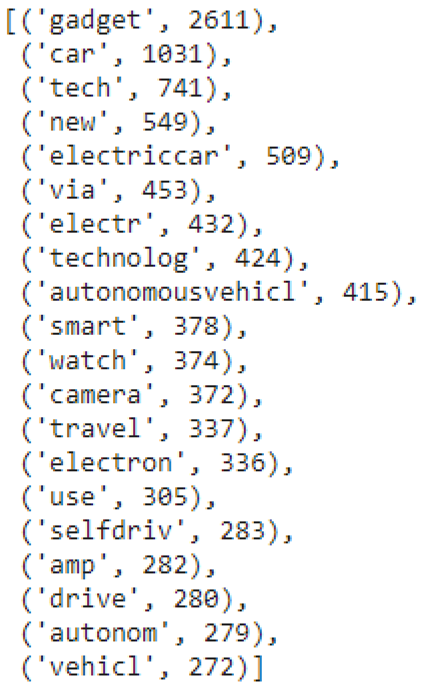
Bag of Words: “It is the representation of text which describes the presence of words within the text data we extract.” “Two similar text fields will contain similar kind of words and therefore have a similar bag of words. From that text we learn about the meaning of the document.”



```
['autobrighttech',
 'autobrighttech',
 'caraccessori',
 'autoaccessori',
 'otoaccessori',
 'vehicleaccessori',
 'dvr',
 'dvrcamera',
 'dvr',
 'dvr',
 'dvr',
 'dvrl',
 'carcamera',
 'carcamerarecord',
 'carcamerajapan',
 'carcamera',
 'carcamera']
```

Figure 8.17 Bag of Words

Most commonly occurring words: These are the most commonly occurring words in the text data.



```
[('gadget', 2611),
 ('car', 1031),
 ('tech', 741),
 ('new', 549),
 ('electriccar', 509),
 ('via', 453),
 ('electr', 432),
 ('technolog', 424),
 ('autonomousvehicl', 415),
 ('smart', 378),
 ('watch', 374),
 ('camera', 372),
 ('travel', 337),
 ('electron', 336),
 ('use', 305),
 ('selfdriv', 283),
 ('amp', 282),
 ('drive', 280),
 ('autonom', 279),
 ('vehicl', 272)]
```

Figure 8.18 Most common words

Word Frequency: “It indicates the number of times each token occurs in a text.”

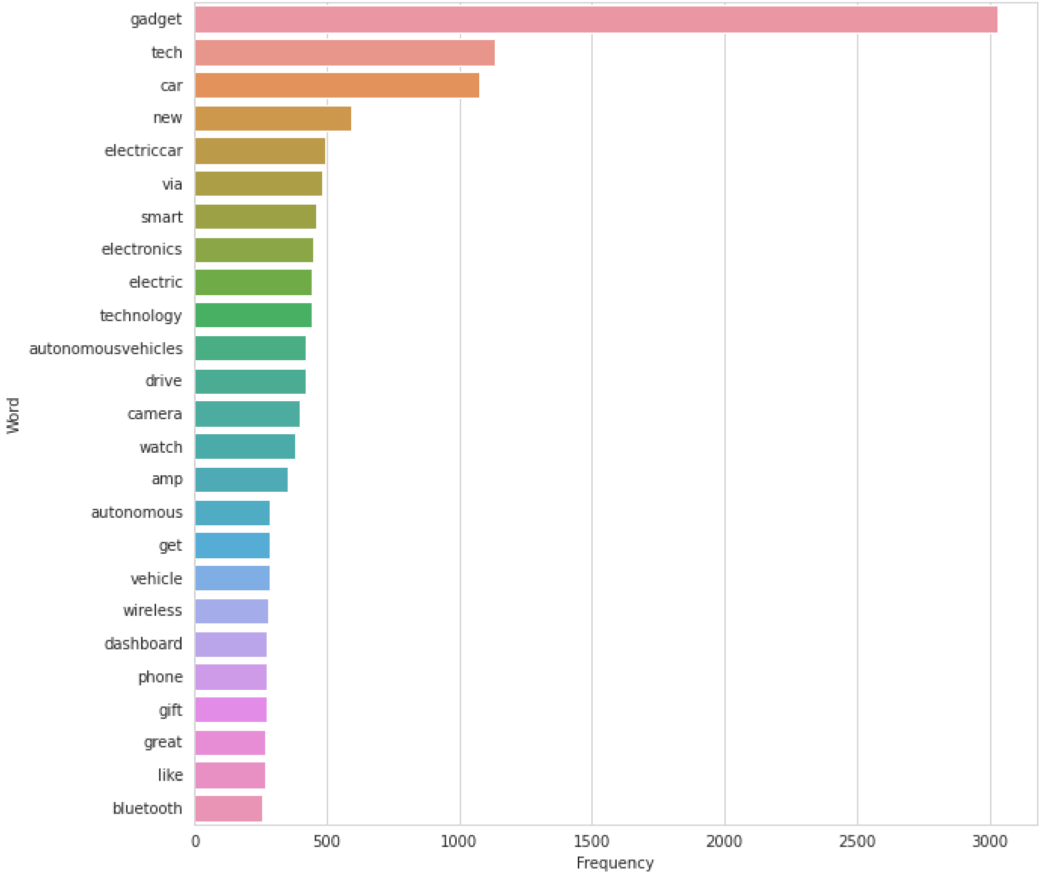


Figure 8.19 Word Frequency

Word Cloud: “Word Cloud is an image composed of words used in a particular text or subject. The size of each word indicates its frequency.” The following figure shows the word cloud for our data”

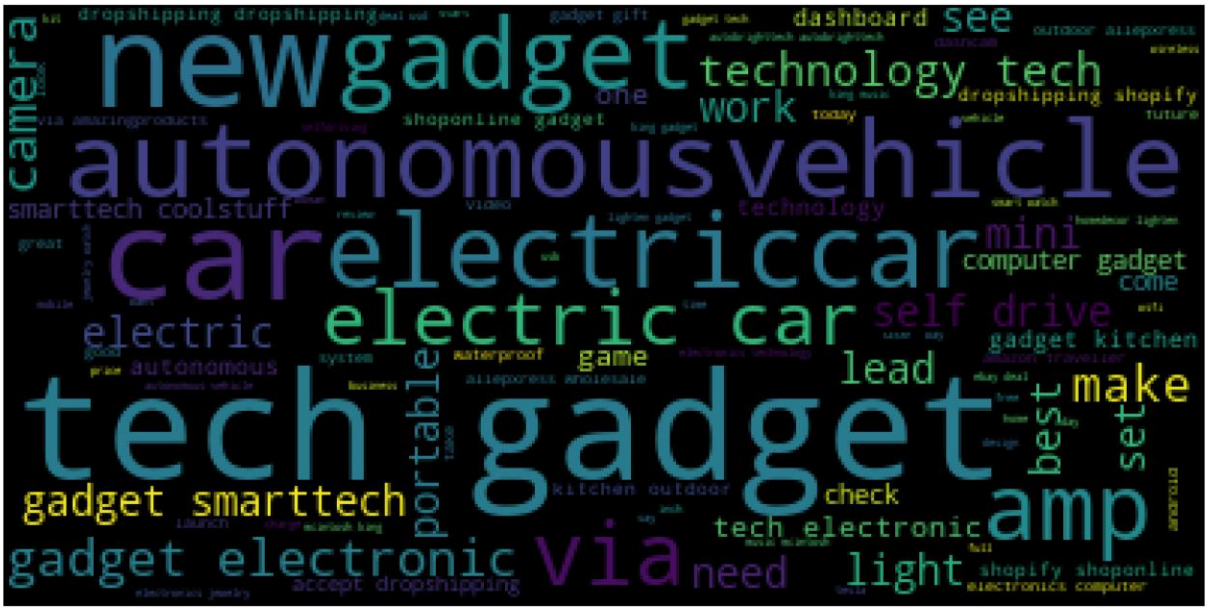


Figure 8.20 Word Cloud

Word frequency distribution: The below figure shows the distribution of word frequency.

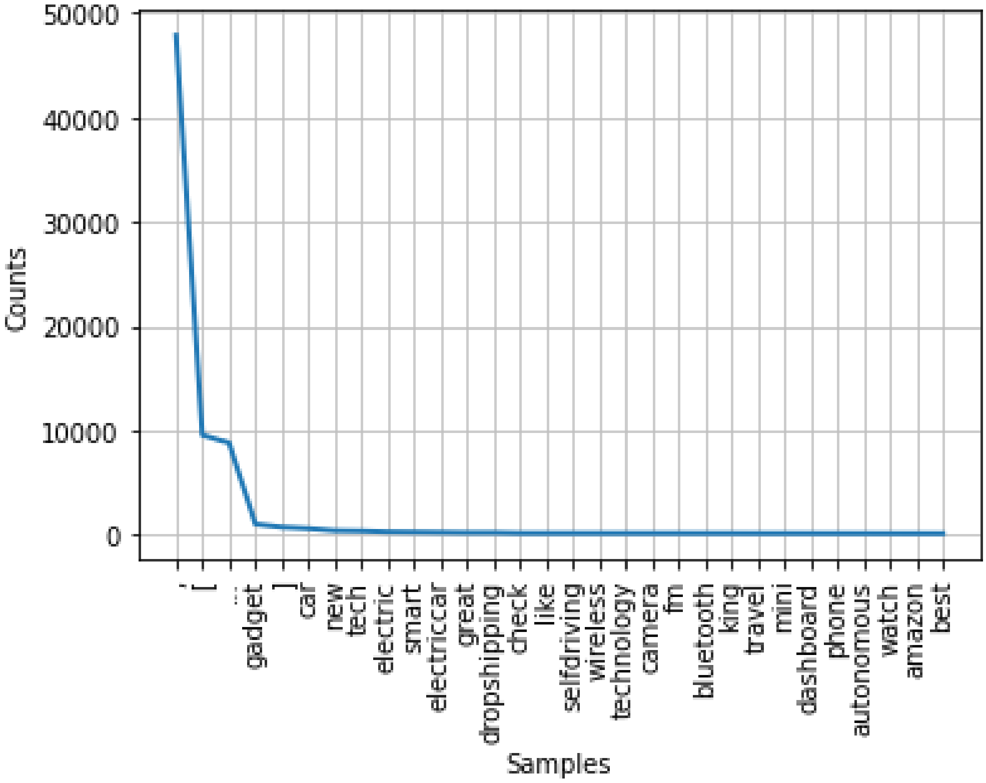
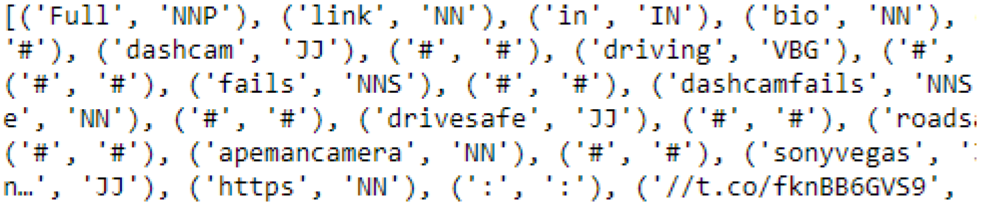


Figure 8.21 Word Frequency Distribution

POS Tagging: “Part-Of-Speech Tagging is the process of assigning different labels knows an POS tags to the words in a sentence.” It tells us about the part of speech of the word.



```
[('Full', 'NNP'), ('link', 'NN'), ('in', 'IN'), ('bio', 'NN'),
'#'), ('dashcam', 'JJ'), ('#', '#'), ('driving', 'VBG'), ('#',
('#', '#'), ('fails', 'NNS'), ('#', '#'), ('dashcamfails', 'NNS
e', 'NN'), ('#', '#'), ('drivesafe', 'JJ'), ('#', '#'), ('roads
('#', '#'), ('apemancamera', 'NN'), ('#', '#'), ('sonyvegas', '
n…', 'JJ'), ('https', 'NN'), (':', ':'), ('//t.co/fknBB6GVS9',
```

Figure 8.22 POS Tagging

Aspects: As per the polarity and sentiment scoring, the aspects will look like following:

	Aspects
0	[projector, clueless]
1	[card, reader, review]
2	[team, email, issue, follow, ups]
3	[picture, 4k, camera, price, adventures, running, camera, performance]
4	[dashcam, camer, cable, ideas, extension]

Table 8.1 Aspects

Chapter 9: Data Modeling

“There are two major approaches in identifying sentiments: lexicon based or rule based and machine learning based. Machine learning based approach uses technique of classification to classify text. Lexicon based approach uses sentiment dictionary to determine with opinion words. It matches them with data to determine the polarity. Lexicon based approach involves calculating the orientation of words or phrases.” (Aung & Myo, 2017)

“We need special preparation before we can start modelling for text data. The words need to be encoded as integers or floating values to use as input to a machine learning algorithm which will be machine readable. It is called feature extraction or vectorization. The text must be parsed to remove words and it is called tokenization.”(Singh & Kaur, 2015)

“The CountVectorizer library of Python provides a simple way to tokenize a collection of text documents and build a vocabulary of known words. It encodes new documents using that vocabulary.” (Saini, 2019)

After we do the vectorising, we go for Classifiers. Classification is a process to recognize, understand and to group the ideas and objects into pre-set categories.it uses the pre-categorized training datasets and the programs use different algorithms to classify future datasets into categories.

“These are classification techniques commonly used: Logistic Regression, Naïve Bayes Classifier, K-Nearest Neighbours, Decision Tree, Random Forest and Support Vector Machines and so on.”

Confusion Matrix: “This explains the performance of a classification model or classifier in a table format. This is used on a set of test data where the true values are known”. Tthese are basic terms:

“True Positive (TP) – model predicted positive and is actually positive”

“True Negative (TN) – model predicted negative and is actually negative”

“False Positive (FP) – model predicted positive but is actually negative”

“False Negative (FN) – model predicted negative but is actually positive” (Mahendran et al., 2020)

Accuracy: “It is a ratio of correctly predicted observation to the total observations.”

“The formula is, Accuracy = TP+TN/TP+FP+FN+TN” (A. & Sonawane, 2016)

Precision: “It is the ratio of correctly predicted positive observations to the total positive observations in prediction”.

“The formula is, Precision = TP/TP+FP” (A. & Sonawane, 2016)

Recall: “It is also called Sensitivity. It is the ratio of correctly predicted observations to the all observations in the actual class.”

“The formula is, Recall = TP/TP+FN” (A. & Sonawane, 2016)

F1 Score: “It takes harmonic mean of Precision and Recall and it is weighted average of them. It is useful to compare the classifiers.”

"The formula is, F1 Score = 2*(Recall * Precision) / (Recall + Precision)" (A. & Sonawane, 2016)

Support: **"**In the specified dataset, Support is the number of actual occurrences of the class. It diagnoses the evaluation process." (A. & Sonawane, 2016)

Macro Average: **"**It is the function to compute F1 score each label. It returns the average without considering the proportion for each label in the dataset." (A. & Sonawane, 2016)

Weighted Average: "It is the function to compute F1 score for each label. It returns the average considering the proportion for each label in the dataset." (A. & Sonawane, 2016)

Lexicon Based Approach:

In the beginning of modeling, we will use Lexicon Based Approach to identify sentiments using AFINN lexicon is the simplest and one of the most important lexicon used for sentiment analysis. "This is done by Finn Årup Nielsen between 2009 and 2011. AFINN-en-165.txt is the current version of lexicon. It contains over 3,300+ words and it has a polarity score associated with each word." (Sarkar, 2019)

Then represents negative score for the user comments with customers expressing their dissatisfaction and positive score for the comments expressing customers' satisfaction and extreme positive score for the customer delight

When we do the EDA for the categorized data, we can see the data spread as this:

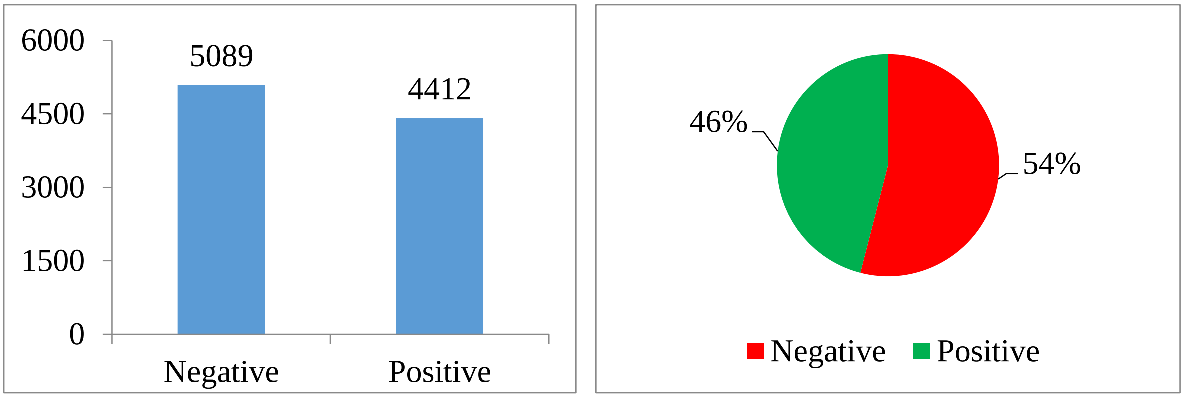


Figure 9.1 and Figure 9.2: EDA of Lexicon

Comparison of Accuracy of Classifiers:

Modeling results for AFINN Lexicon: In this approach, we have used SVC, Decision Tree, Random Forest, Gradient Boosting and Gaussian Naive Bayes classifiers here. With these classifiers, Random Forest gave the highest accuracy of 84% and SVC gave lowest accuracy of 55% as shown below:

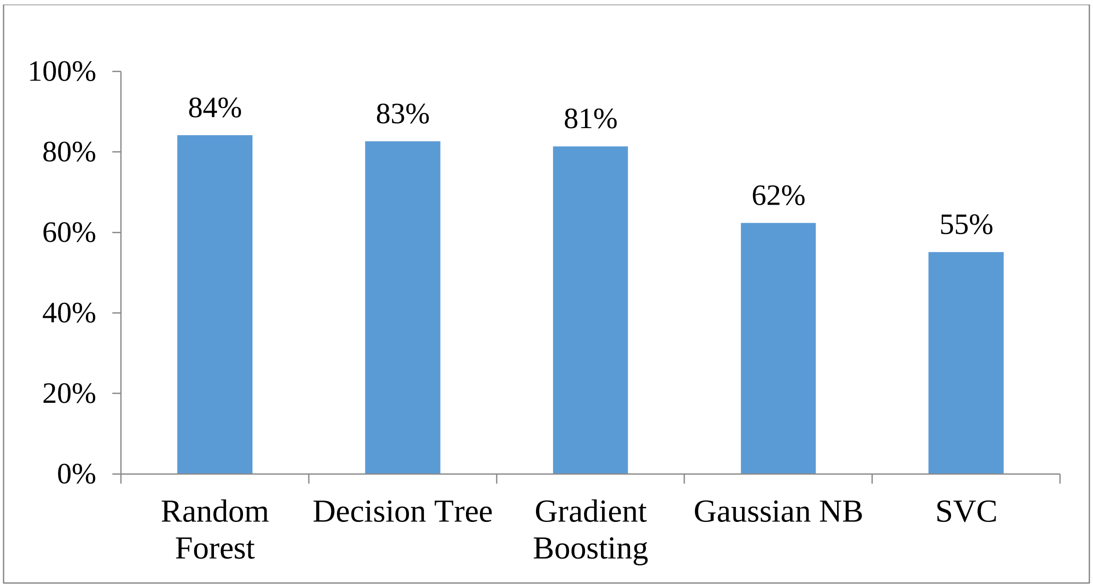


Figure 9.3 Comparison of Accuracies - Lexicon

The result for Random Forest classifier is shown below:

Random Forest	Precision	Recall	F1-Score	Support
Negative	0.91	0.81	0.86	1422
Positive	0.76	0.88	0.82	963
Accuracy			0.84	2385
Macro Average	0.84	0.85	0.84	2385
Weighted Average	0.85	0.84	0.84	2385

Table 9.1 Random Forest Result – Lexicon

True Label	Negative	1155	267
	Positive	111	852
		Negative	Positive
		Predicted Label	

Table 9.2 Random Forest Confusion Matrix - Lexicon

As it is understood from the above results and confusion matrix, it is not optimistic to say Lexicon model predicting accurately. There are several reasons as disadvantages for the Lexicon model using AFINN dictionary.

Lexical approach can be a quick way to pick up the phrases but it has some limitations. It doesn’t give much creativity. The major limitation is, it incorrectly scores the sentiment of opinion words using existing Lexicons.(Asghar et al., 2017)

TextBlob Sentiment Analysis: In Machine Learning approach, we will use TextBlob package from Python for sentiment scoring. It performs the sentiment analysis by calculating the polarity of the user comments from the dataset from -1 to 1. It uses NLTK corpora. The Facebook/Twitter user comments dataset is trained using different classifiers. (Mittal & Patidar, 2019)
TextBlob has a large number of corpora set. It provides many stemmers and also many algorithms which help to perform text analysis. (Kumar, 2019)

The EDA of TextBlob Polarity:

The data sample polarity by TextBlob:

Polarity	User Comments	Category
-0.100000	gadget wide angl macro phone camera lens pc set	Negative
0.295455	new small dreamcatch perfect hang rearview mir...	Positive
0.5	sturdi unit otherwis ok	Positive
0.7	look good	Positive
-0.283333	increas traffic volum creat problem citi aroun..	Negative

Table 9.3 Polarity by TextBlob

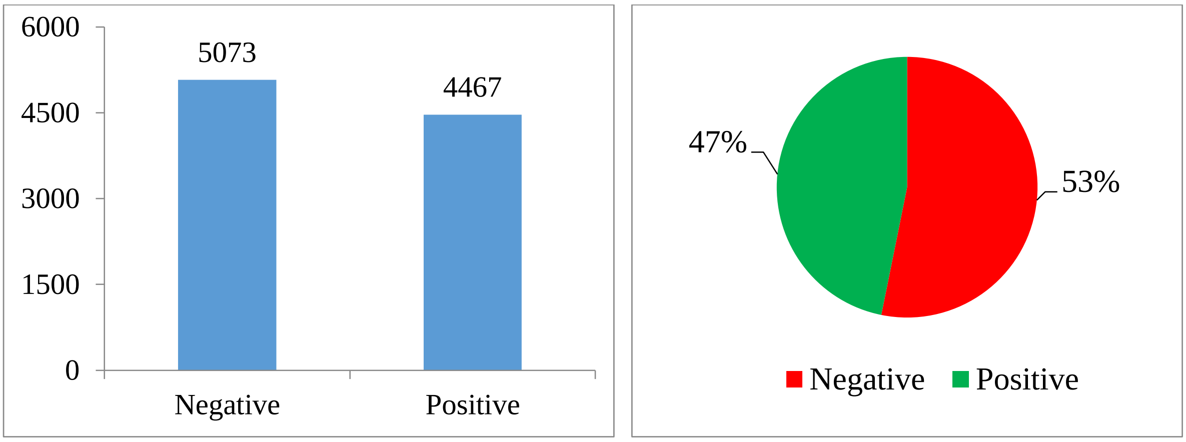


Figure 9.4 and Figure 9.5 EDA of TextBlob

As it is observed from the above charts, the scoring result is almost similar to Lexicon approach.

The comparison of classifiers: Random Forest has given the accuracy of 84% similar to Lexicon approach and SVC has the lowest accuracy of 54%.

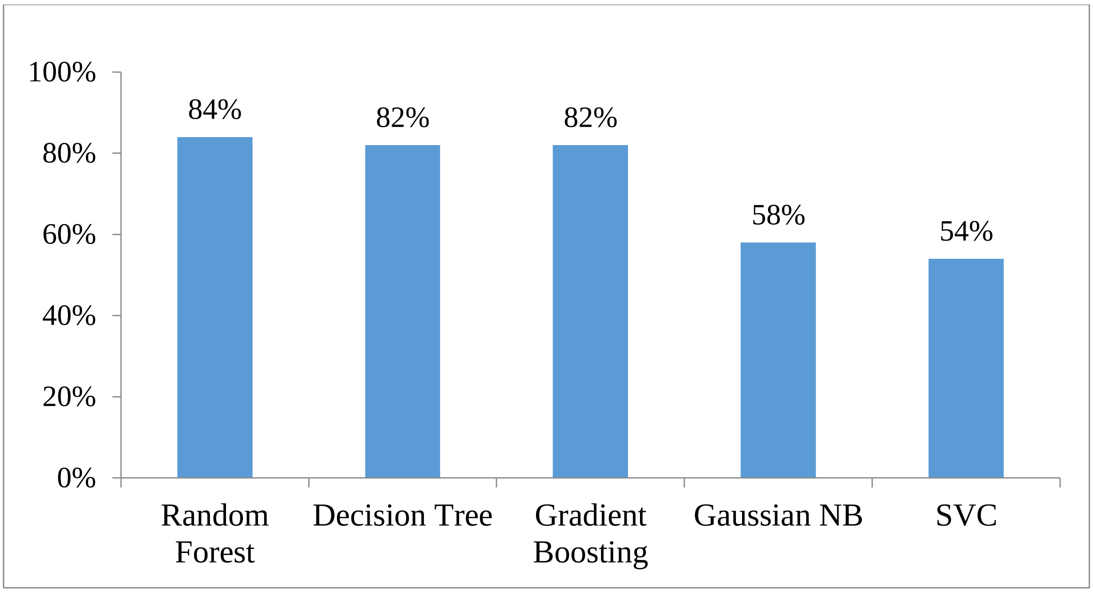


Figure 9.6 Comparison of Accuracies - TextBlob

The model result of Random Forest Classifier is shown below:

Random Forest	Precision	Recall	F1-Score	Support
Negative	0.93	0.81	0.86	1462
Positive	0.75	0.90	0.82	923
Accuracy			0.84	2385
Macro Average	0.84	0.85	0.84	2385
Weighted Average	0.86	0.84	0.85	2385

Table 9.4 Random Forest Result – TextBlob

		Predicted Label	
		Negative	Positive
True Label	Negative	1180	282
	Positive	90	833

Table 9.5 Random Forest Confusion Matrix - TextBlob

From the modeling results and confusion matrix, the results obtained from AFINN Lexicon and TextBlob approaches are similar and Random Forest being giving highest accuracy of 84%

We will continue with another approach, it is called VADER sentiment analysis.

VADER Sentiment Analysis: “VADER is acronym of Valence Aware Dictionary and sEntiment Reasoner”. It is a Lexicon based approach. It comes under Rule-based systems. It is sensitive to the polarity and intensity of the emotions of the user. Due to the advantages it provides, it has been used more in predictive analysis. It is not necessary to have training data as it can be constructed from a valence based and able to generalize sentiment Lexicon. (DEO, 2020)

The EDA of VADER Sentiment:

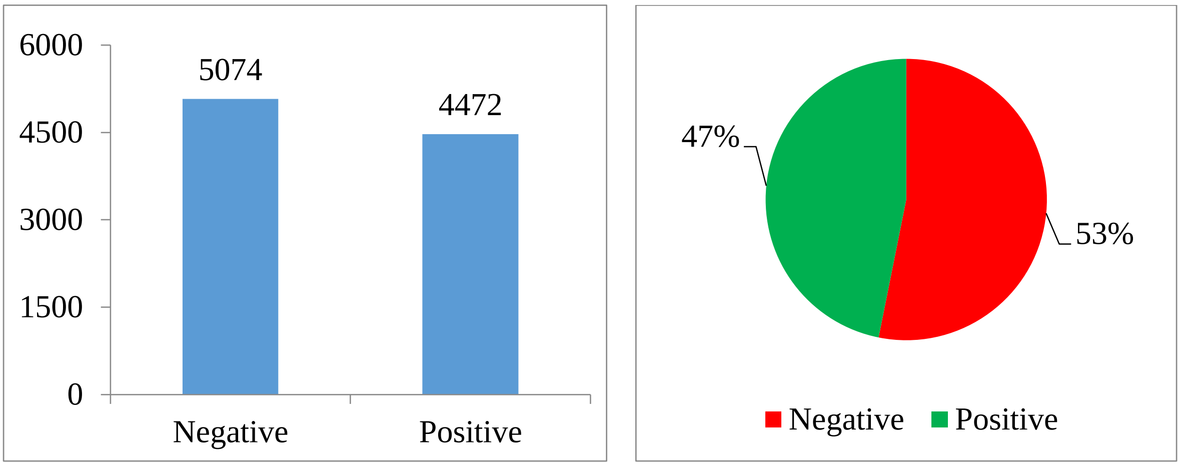


Figure 9.7 and Figure 9.8 EDA of VADER

The modelling results with classifier accuracy shown in below graph:

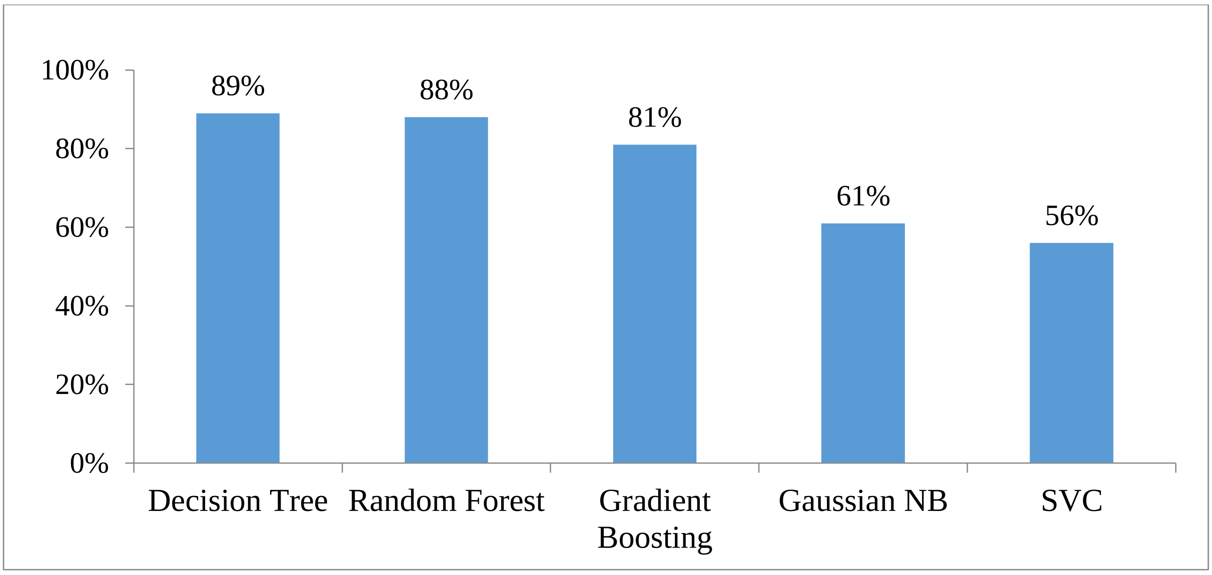


Figure 9.9 Comparison of Accuracies – VADER

Decision Tree	Precision	Recall	F1-Score	Support
Negative	0.92	0.89	0.90	1345
Positive	0.86	0.90	0.88	1042
Accuracy			0.89	2387
Macro Average	0.89	0.89	0.89	2387
Weighted Average	0.89	0.89	0.89	2387

Table 9.6 Decision Tree Result – VADER

True Label	Negative	1192	153
	Positive	109	933
		Negative	Positive
		Predicted Label	

Table 9.7 Decision Tree Confusion Matrix - VADER

As shown above results of VADER sentiment analysis, Decision Tree classifier gives highest accuracy of 89% and SVC with lowest of 56%.

TF-IDF Vectorizer: “TF stands for Term Frequency and IDF is Inverse Document Frequency.”
“This is an algorithm to text into meaningful representation of numbers. It is used to fit predictive machine algorithm.”
“It converts the text documents into vector models. It is based on the occurrence of words in the documents without taking considering the exact ordering.”

“TF measures how frequently a term occurs in a document and IDF measures how importance the term is.”

The formula of TF:

“TF = Number of times term appears in a document
Total number of terms in the document”

The formula of IDF:

“IDF = log_e of Total number of documents
Number of documents with term in it”

When the same modeling checked with IF-IDF approach, we got the following modeling results:

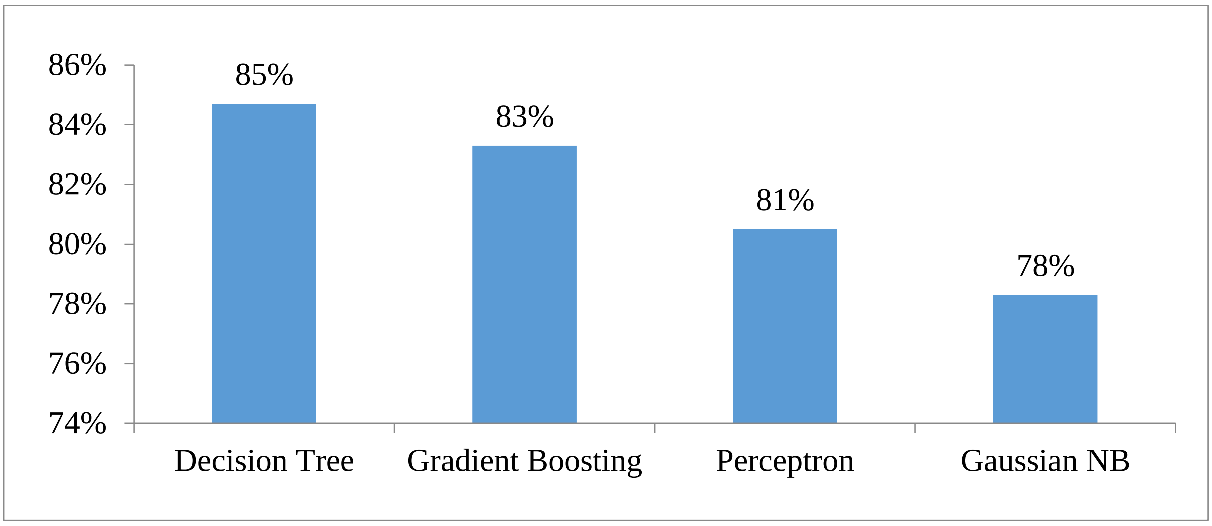


Figure 9.10 Comparison of Accuracies – TF-IDF

Decision Tree is giving the highest accuracy of 85% in IF-IDF method.

Word Embedding: It is a new way of representing words into vectors. It will redefine the high dimensional word features into low dimensional feature, however it preserves the contextual similarity in the corpus. The Word Embedding are widely used in Deep Learning models especially in CNN and RNN.
Word2Vec and GloVe are the two major models used to create word embedding of a text.

Word2Vec: The input for Word2Vec is a text corpus. It produces the word vectors as output.
"The vocabulary is constructed first from the training text data. It then learns vector representation of words. The resulting word vector will be used as features in NLP applications." (Google, 2013)
GloVe: "GloVe is an unsupervised learning algorithm used to obtain vector representation for words. The training of algorithm is conducted by performing on aggregated global word-word co-occurrence statistics from a corpus. (Jeffrey Pennington, Richard Socher, 2014)
We have used Word2Vec in this project." (Thi-Luong et al., 2017)

Supervised Learning: "It is a machine learning method where we train the machine using data with labelled and it means some data is tagged with correct answer. It is like in the presence of a supervisor. Classification and Regression are Supervised Learning techniques." (Kawade & Oza, 2017)

Unsupervised Learning: "It is a machine learning method where we train using information that is neither classified nor labelled. It allows algorithm to act on the data without guidance or supervision. Clustering and Association are unsupervised techniques." (Kawade & Oza, 2017)

In Word2Vec approach, we have used sequential modeling with embedding, bidirectional and dense with sigmoid activation.
"LSTM is an artificial RNN architecture used in Deep Learning, "GRU is gating mechanism in RNNs" (Xue & Li, 2018)
With this, the model analysis showed results as shown below:

Word2Vec	Precision	Recall	F1-Score	Support
Negative	0.59	0.67	0.63	1022
Positive	0.53	0.44	0.48	846
Accuracy			0.57	
Macro Average	0.56	0.56	0.56	1868
Weighted Average	0.56	0.57	0.56	1868

Table 9.8 Word2Vec Result

The Confusion Matrix for Word2Vec is:

True Label	Negative	687	335
	Positive	471	375
		Negative	Positive
		Predicted Label	

Table 9.9 Confusion Matrix - Word2Vec

As it is visible with the result and confusion matrix, Word2Vec has not given much better result in accuracy. It has given the accuracy of 57% and the Precision and Recall are also not showing good result.

Transfer Learning: "It acquires the knowledge while solving one problem and applies it to a different but related problem. It re-uses the pre-trained model on a new problem. It has become popular now in Deep learning due to its ability to train deep neural networks with comparatively little data. The Deep Learning model is trained don large dataset and used to perform similar tasks on another dataset. These are called pre-trained models. It is better to use a pre-trained model as a starting point instead of building model from scratch." (Zhang et al., 2011)
The transformer model was introduced by Google.

We have used fine-tune BERT for user comments classification here. The model is fine-tuned a BERT model to perform text classification, it is done with the help of Transformers library.

Fine Tuning with BERT: Google has open-sourced a new technique in 2018 called BERT, with this anyone can train their models as transfer learning “It is the first deeply bidirectional unsupervised language representation. It is pre-trained using plain text corpus.”

The following table shows the result of fine tuning with BERT:

Fine Tuning BERT	Precision	Recall	F1-Score	Support
Negative	0.79	0.57	0.67	969
Positive	0.43	0.68	0.53	463
Accuracy			0.61	1432
Macro Average	0.61	0.63	0.6	1432
Weighted Average	0.68	0.61	0.62	1432

Table 9.10 Fine Tuning BERT Result

The Confusion Matrix from fine tuning BERT is as shown below:

True Label	Negative	557	412
	Positive	146	317
		Negative	Positive
		Predicted Label	

Table 9.11 Confusion Matrix – Fine Tuning with BERT

The accuracy in fine tuning with BERT is 61% better than Word2Vec but still we need to improve the model for higher accuracy.

Chapter 10: Data Evaluation

As we can see, we have seen various approaches in solving our problem of identifying voice of customers for automotive gadgets from Facebook/Twitter user comments. The significant approaches we used are Lexicon with AFINN vocabulary, classifiers with TextBlob Polarity, VADER Sentiment, TF-IDF Vectoriser, Word Embedding with Word2Vec and Transfer Learning using fine-tuned BERT model and compared.

The primary approach here was NLP. Our dataset consists of 9,546 rows with user comments extracted from Facebook and twitter APIs. We started with our approach with Lexicon using AFINN vocabulary, TextBlob polarity followed by VADER sentiment scoring.

To convert text to features, we have used text to features like CountVectorizor, TF-IDF, Word2Vec.

The initial data had to clean as it had many noises such spelling mistakes, handles, emojis, case sensitive texts and so on. Data cleaning was an important phase where we need to balance the dataset as sometimes the major class contributes more data after polarity or sentiment scoring which makes model to give wrong or illusion results.

Here, we have used Lexicon scoring, TextBlob polarity, VADER sentiment scoring for the labelling of category of user comments sentiments.

We also used Deep Learning techniques like Word2Vec and Transfer learning using fine tuning with BERT which are latest techniques in the field.

For classifiers, we mainly used SVC, Random Forest, Decision Tree, Gaussian NB and Gradient Boosting classifiers.

As observed the performance of classifiers and deep learning models, it is important to have accurate labelling for all comments.

S.N.	Approach	Accuracy	Classifier / Model with best result
1	VADER Sentiment	89%	Decision Tree
2	Word Embedding - TF-IDF Vectorizer	85%	Decision Tree
3	Lexicon Vocabulary	84%	Random Forest
4	TextBlob Polarity	84%	Random Forest
5	Transfer Learning - Fine Tuning with BERT	61%	Transfer Learning Fine Tuning
6	Word Embedding - Word2Vec	57%	Sequential-LSTM, GRU, Dense

Table 10.1 Accuracies of Modelling approaches

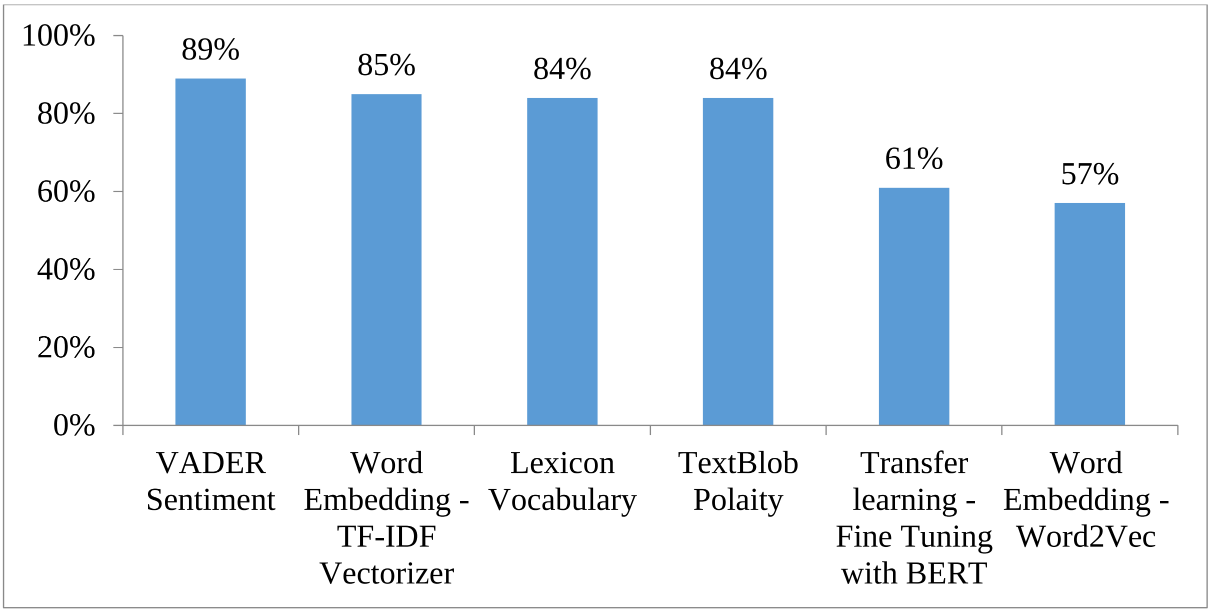


Figure 10.1 Comparison of accuracies

As we can see, VADER sentiment scoring with Decision Tree Classifier has given the highest accuracy of 89% for the sentiment analysis.

There are various factors affecting model performance for other approaches especially in deep learning techniques for which we need to have more optimized data.

Chapter 11: Deployment

As an implementation for deployment, the following image is illustration of the dashboard proposed as future work.

The data pipeline shows the deployment plan for future as business requirement to develop as front end API as an application.

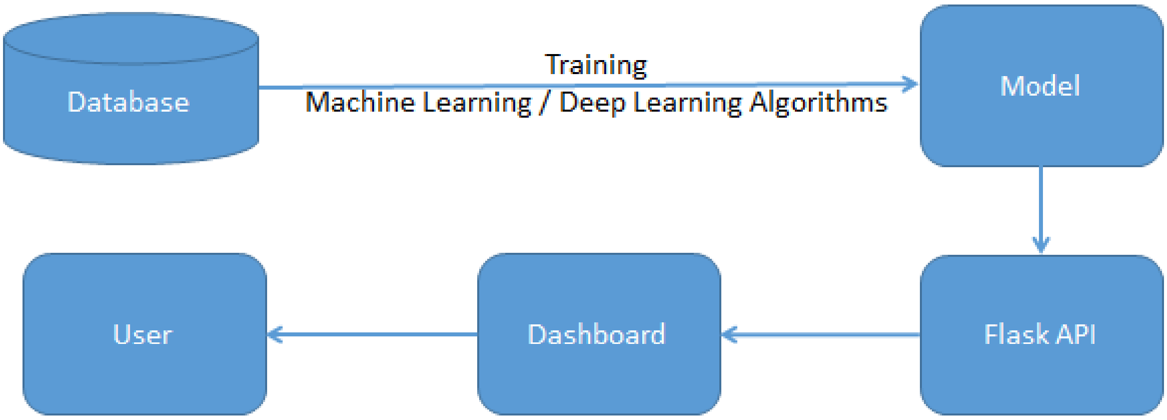
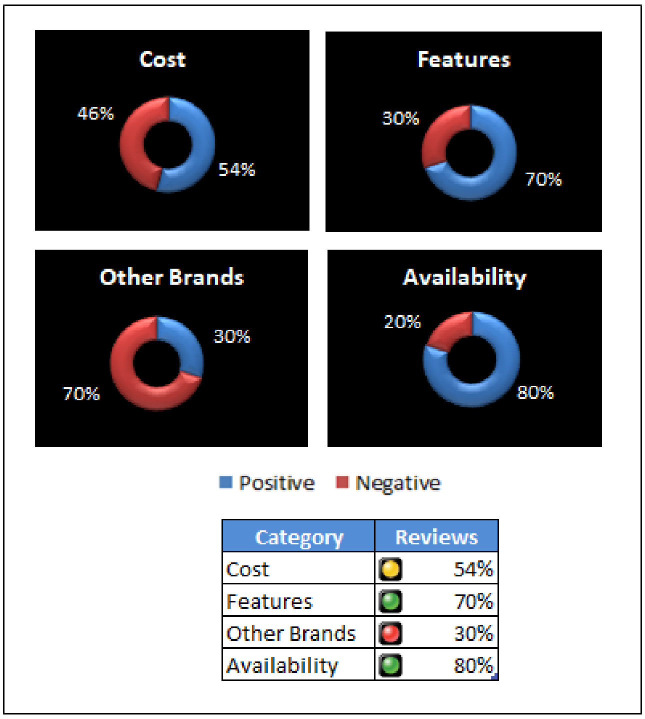


Figure 11.1 Deployment Proposal

This is a proposal as future work where the dashboard gives the output from API from the machine/deep learning algorithms for multi label features with end to end plug in API.



Category	Reviews	
Cost	🟡	54%
Features	🟢	70%
Other Brands	🔴	30%
Availability	🟢	80%

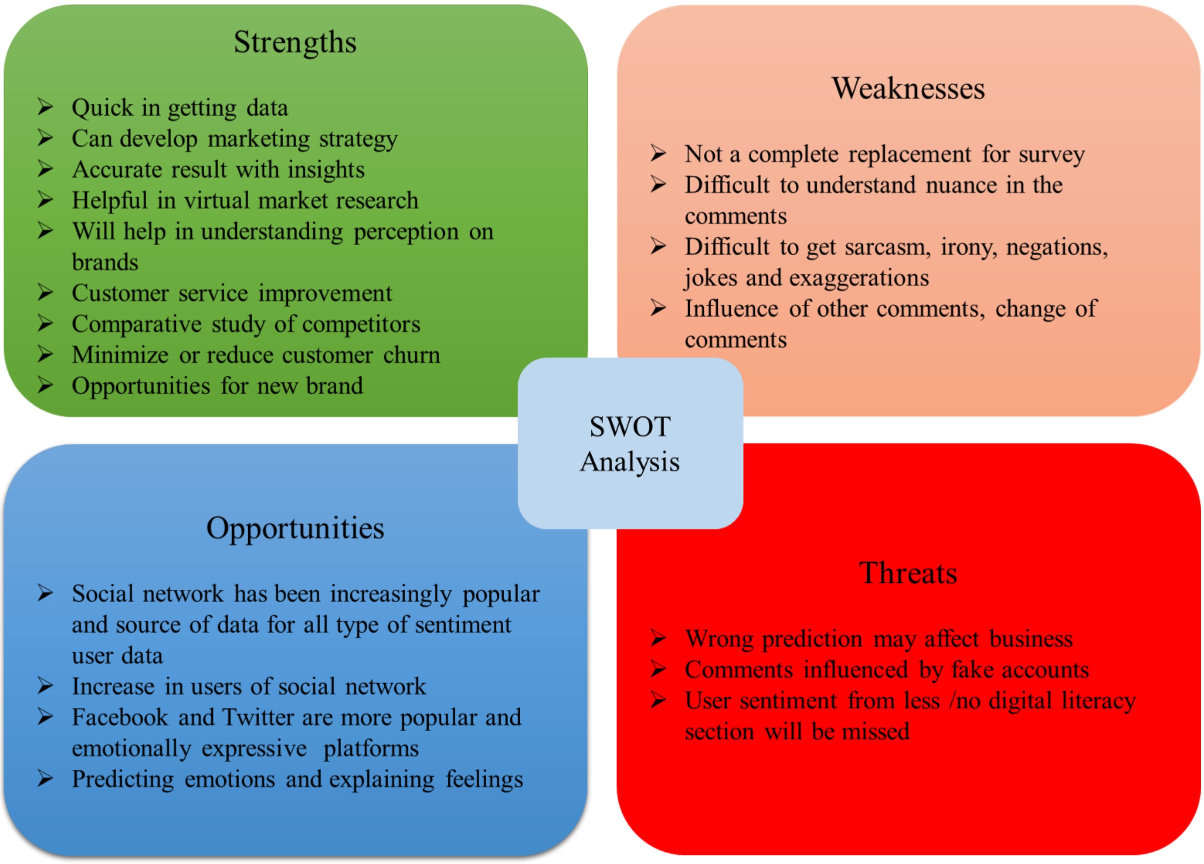
Figure 11.2 Illustration of Dashboard

Chapter 12: Analysis and Results

As mentioned in the Data Evaluation chapter, we got an accuracy of 89% with Decision Tree classifier for VADER sentiment scoring. There could be various factors affecting the performance of other classifiers, deep learning and transfer learning with fine tune BERT models. The model performance can be improved with following fine tuning processes:

- Having more dataset
- Topic modeling for more data points
- Handling class imbalance for bigger dataset
- Domain specific support for accurate labelling by specific corpus
- More efficient negation handling
- Exploring other techniques like GloVe, fastText for word embedding and comparing the results with the results obtained in this study

SWOT Analysis:



Strengths

- ➢ Quick in getting data
- ➢ Can develop marketing strategy
- ➢ Accurate result with insights
- ➢ Helpful in virtual market research
- ➢ Will help in understanding perception on brands
- ➢ Customer service improvement
- ➢ Comparative study of competitors
- ➢ Minimize or reduce customer churn
- ➢ Opportunities for new brand

Weaknesses

- ➢ Not a complete replacement for survey
- ➢ Difficult to understand nuance in the comments
- ➢ Difficult to get sarcasm, irony, negations, jokes and exaggerations
- ➢ Influence of other comments, change of comments

SWOT Analysis

Opportunities

- ➢ Social network has been increasingly popular and source of data for all type of sentiment user data
- ➢ Increase in users of social network
- ➢ Facebook and Twitter are more popular and emotionally expressive platforms
- ➢ Predicting emotions and explaining feelings

Threats

- ➢ Wrong prediction may affect business
- ➢ Comments influenced by fake accounts
- ➢ User sentiment from less /no digital literacy section will be missed

Figure 12.1 SWOT Analysis

Chapter 13: Conclusions and Recommendations for future work

In this project, we have tried identifying voice of customers for automotive gadgets using Facebook and Twitter user comments. We have used various approaches of data science to extract and study sentiments of users.

As mentioned in the Data Evaluation and Analysis and Results, machine learning and deep learning algorithms are able to do identify the voice of customers for automotive gadgets for the data extracted from Facebook and Twitter.
We can improve the modelling results by having some more data treatment and with considerable amount of excess balanced data.

As a future work, it is planned to create a dashboard as mentioned in the previous chapter and to build a state of the art front end API using Flask API, which serves the requirement of enterprises to get the virtual market survey, to get voice of customers and for market research. As an additional scope of future work is to detect emotions also from the user comments, this will also help to predict the user comments. The additional work will also involve to identify comments from fake accounts which is a threat and to find solutions.

Bibliography

A., V., & Sonawane, S. S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications, 139*(11), 5–15. https://doi.org/10.5120/ijca2016908625

Asghar, M. Z., Khan, A., Ahmad, S., Qasim, M., & Khan, I. A. (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. *PLoS ONE*, *12*(2), 1–22. https://doi.org/10.1371/journal.pone.0171649

Aung, K. Z., & Myo, N. N. (2017). Sentiment analysis of students' comment using lexicon based approach. *Proceedings - 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017*, 149–154. https://doi.org/10.1109/ICIS.2017.7959985

Azizah, N., Suyadi, I., & Kusumawati, A. (2017). the Development Perceived of Product Comments Trust-Ability on Social Media: a Social Information Processing (Sip) Theory Approach (Survey on Facebook Community of Taiwanese Students Colleges). *Russian Journal of Agricultural and Socio-Economic Sciences*, *62*(2), 129–137. https://doi.org/10.18551/rjoas.2017-02.15

Benson, E., Haghighi, A., & Barzilay, R. (2011). Event discovery in social media feeds. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 389–398.

Bilyk, V. (2017). *WHY BUSINESS APPLIES SENTIMENT ANALYSIS? 5 SUCCESSFUL EXAMPLES*. https://theappsolutions.com/blog/development/sentiment-analysis-for-business/

DEO, G. (2020). Predictive Analysis of Resource Usage Data in Academic Libraries using the VADER Sentiment Algorithm. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3663332

Google. (2013). *word2vec*. https://code.google.com/archive/p/word2vec/

Hu, G., Bhargava, P., Fuhrmann, S., Ellinger, S., & Spasojevic, N. (2017). Analyzing Users' Sentiment Towards Popular Consumer Industries and Brands on Twitter. *IEEE International Conference on Data Mining Workshops, ICDMW*, *2017-Novem*, 381–388. https://doi.org/10.1109/ICDMW.2017.55

Jeffrey Pennington, Richard Socher, C. D. M. (2014). *GloVe: Global Vectors for Word*

Representation. https://nlp.stanford.edu/projects/glove/

Kawade, D. R., & Oza, D. K. S. (2017). Sentiment Analysis: Machine Learning Approach. *International Journal of Engineering and Technology*, *9*(3), 2183–2186. https://doi.org/10.21817/ijet/2017/v9i3/1709030151

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2018). Natural Language Processing : State of The Art , Current Trends and Challenges Natural Language Processing : State of The Art , Current Trends and Challenges Department of Computer Science and Engineering Manav Rachna International University , Faridabad-. *ArXiv Preprint ArXiv*, *August 2017*.

Kumar, E. (2019). *A Real-Time Twitter Sentiment Analysis and Visualization System : TwiSent*. *June*, 3323–3330.

Mahendran, N., Durai Raj Vincent, P. M., Srinivasan, K., Sharma, V., & Jayakody, D. K. (2020). Realizing a Stacking Generalization Model to Improve the Prediction Accuracy of Major Depressive Disorder in Adults. *IEEE Access*, *8*(March), 49509–49522. https://doi.org/10.1109/ACCESS.2020.2977887

Mittal, A., & Patidar, S. (2019). Sentiment analysis on twitter data: A survey. *ACM International Conference Proceeding Series*, 91–95. https://doi.org/10.1145/3348445.3348466

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 1320–1326. https://doi.org/10.17148/ijarcce.2016.51274

Pawar, K. K., Shrishrimal, P. P., & Deshmukh, R. R. (2015). *researchpaper-Twitter-Sentiment-Analysis-A-Review*. *6*(4), 957–964.

Ramsaran-Fowdar, R. R. (2013). The Implications of Facebook Marketing for Organizations. *Contemporary Management Research*, *9*(1), 73–84. https://doi.org/10.7903/cmr.9710

Saini, A. (2019). Anuj@IEEE BigData 2019: A Novel Code-Switching Behavior Analysis in Social Media Discussions Natural Language Processing. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 5957–5961. https://doi.org/10.1109/BigData47090.2019.9005961

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: Facebook and Twitter perspectives. *Advances in Science, Technology and Engineering Systems*, *2*(1), 127–133. https://doi.org/10.25046/aj020115

Sankaran, K. (2017). *Breaking through text clutter with natural language processing*.

https://www.latentview.com/blog/breaking-text-clutter-natural-language-processing/

Sarkar, D. (2019). Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. *Text Analytics with Python*, 1–674. https://dl.acm.org/citation.cfm?id=3360099

Sharma, S., Srinivas, P., & Balabantaray, R. C. (2013). Emotion Detection using Online Machine Learning Method and TLBO on Mixed Script. *In Proceedings of Language Resources and Evaluation Conference,* 47.53. https://interop2016.github.io/pdf/INTEROP-11.pdf

Shukri, S. E., Yaghi, R. I., Aljarah, I., & 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 2015*, *November*. https://doi.org/10.1109/AEECT.2015.7360594

Singh, R., & Kaur, R. (2015). Sentiment Analysis on Social Media and Online Review. *International Journal of Computer Applications, 121*(20), 44–48. https://doi.org/10.5120/21660-5072

Sinha, A. K. (2018). *Digital India*. *November,* 170–184. https://doi.org/10.4018/978-1-5225-3787-8.ch011

Thi-Luong, N., My-Linh, H., Phuong, L.-H., & Thi-Minh-Huyen, N. (2017). *Using Distributed Word Representations in Graph-Based Dependency Parsing for Vietnamese*. https://doi.org/10.15625/vap.2016.00098

Wikipedia. (2020). *Cross-industry standard process for data mining*. https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Win, T., Thu, P., & Tun, Z. A. R. (2017). Sentiment orientation system of automotive reviews using multinomial naive bayes classifier at document level. *International Journal of Advances in Electronics and Computer Science, 7*, 11–15.

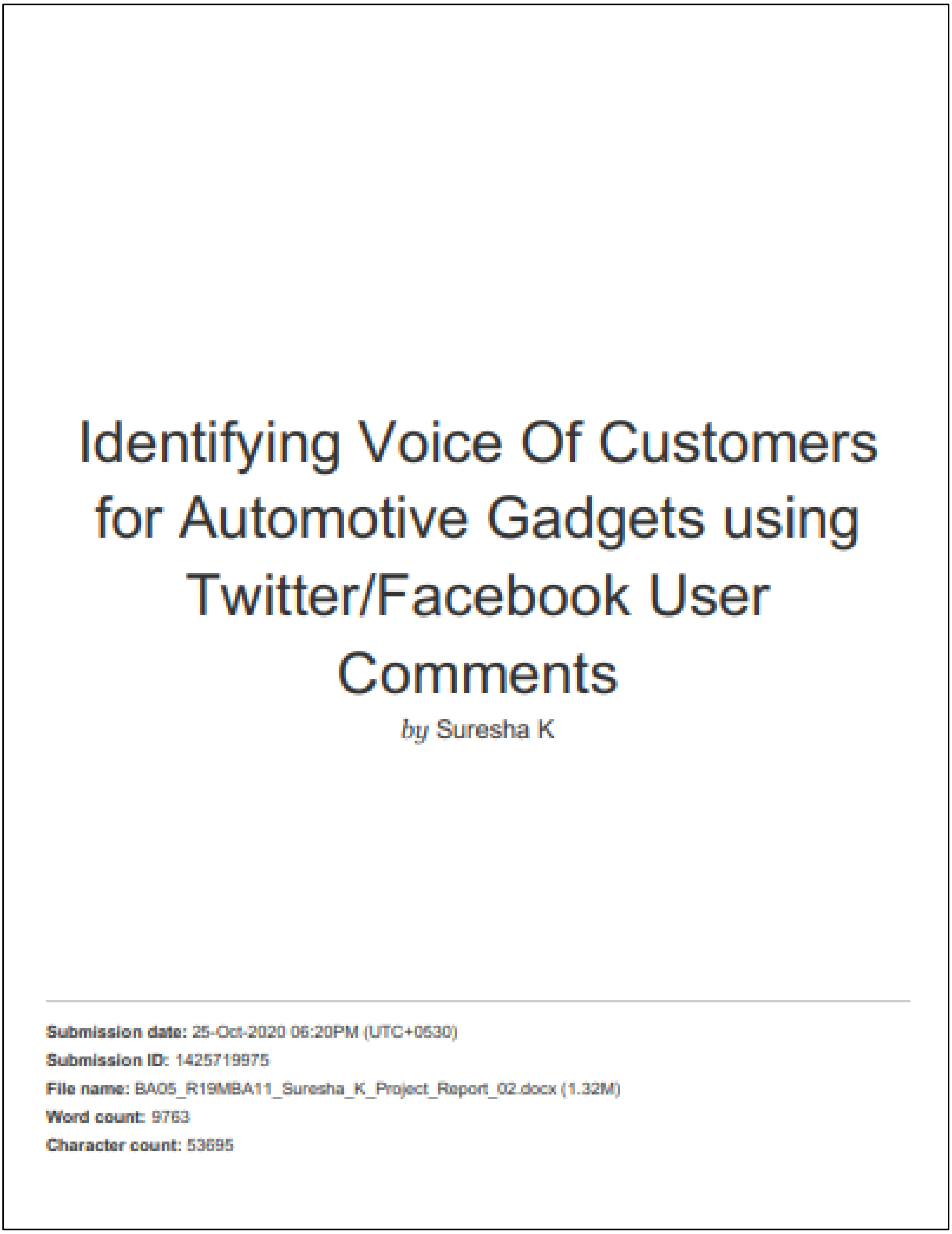
Xue, W., & Li, T. (2018). Aspect based sentiment analysis with gated convolutional networks. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 1,* 2514–2523. https://doi.org/10.18653/v1/p18-1234

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. *Proceedings - IEEE International Conference on Data Mining, ICDM,* 427–434. https://doi.org/10.1109/icdm.2003.1250949

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories Technical Report*, *89*.

Appendix

Plagiarism Report[1]



Identifying Voice Of Customers for Automotive Gadgets using Twitter/Facebook User Comments

by Suresha K

Submission date: 25-Oct-2020 06:20PM (UTC+0530)

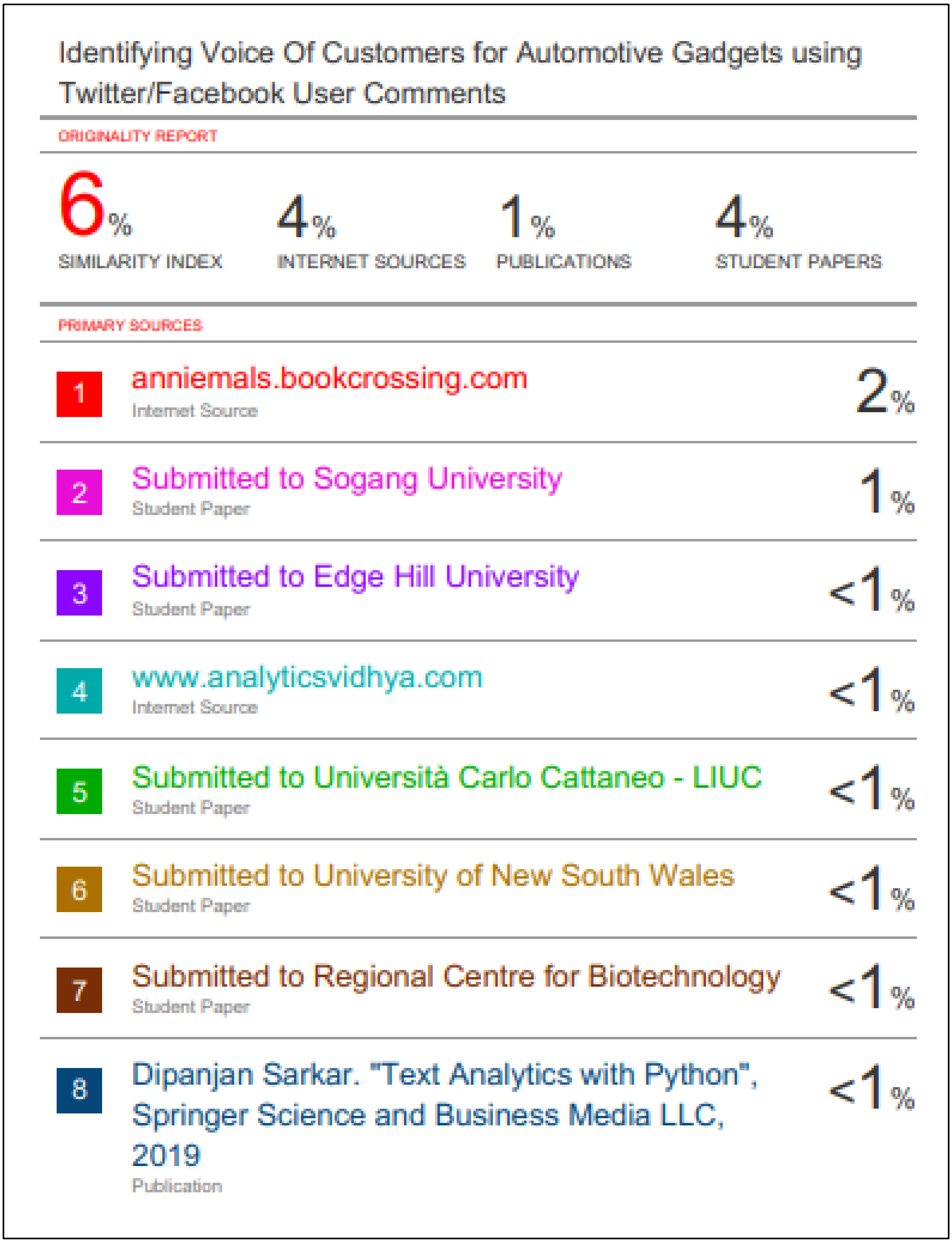
Submission ID: 1425719975

File name: BA05_R19MBA11_Suresha_K_Project_Report_02.docx (1.32M)

Word count: 9763

Character count: 53695

[1] Turnitn report to be attached from the University.



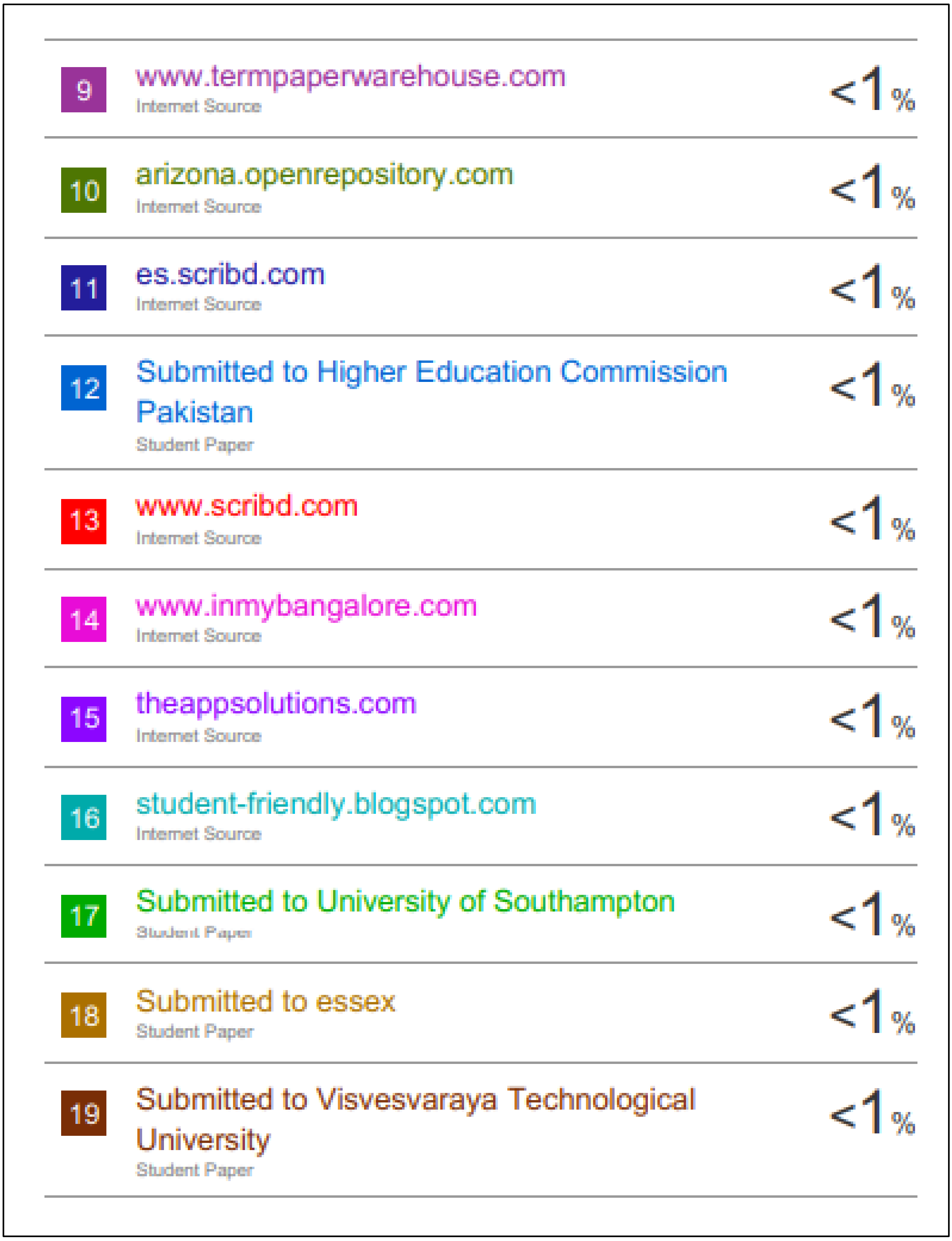
Identifying Voice Of Customers for Automotive Gadgets using Twitter/Facebook User Comments

ORIGINALITY REPORT

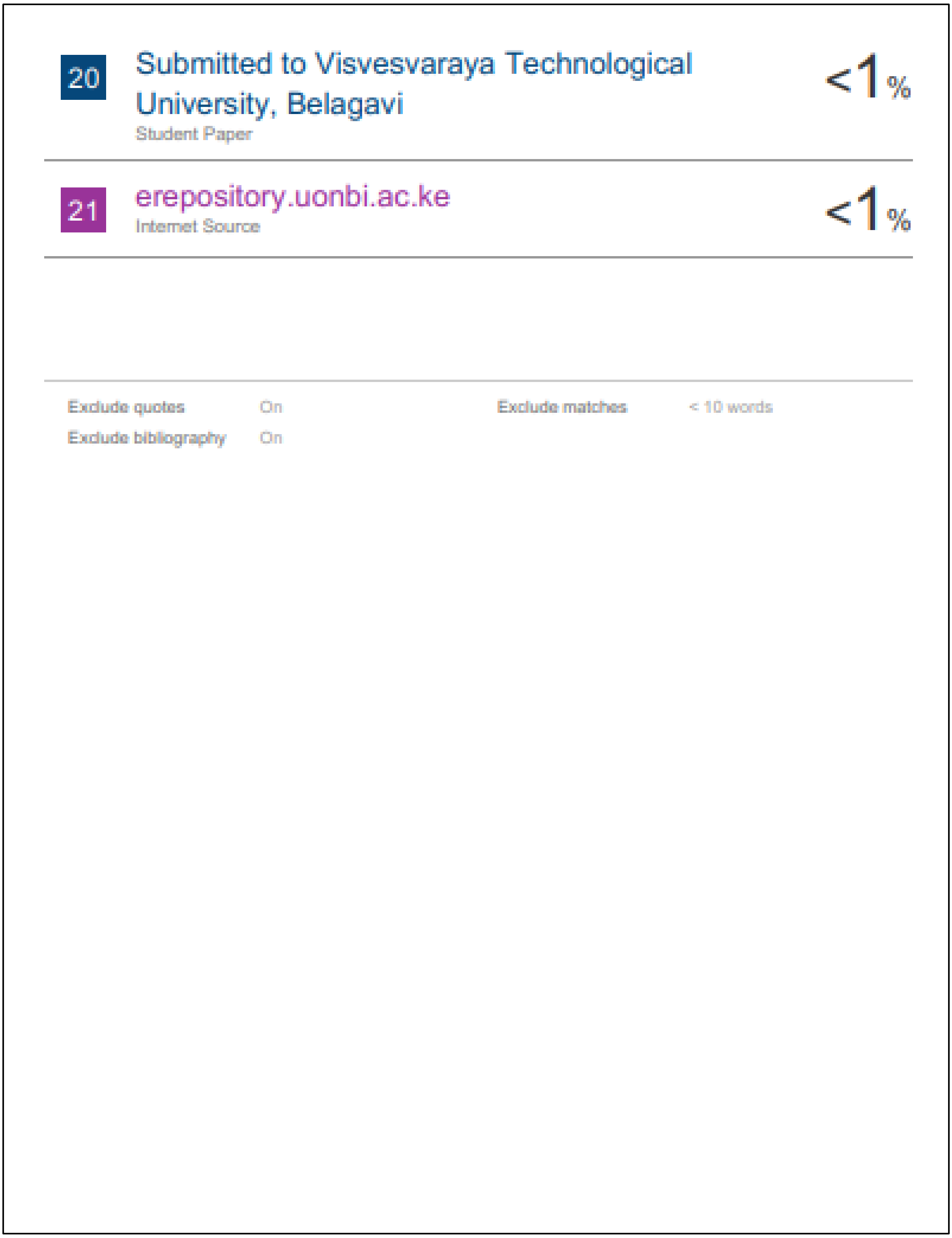
6%	4%	1%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	anniemals.bookcrossing.com Internet Source	2%
2	Submitted to Sogang University Student Paper	1%
3	Submitted to Edge Hill University Student Paper	<1%
4	www.analyticsvidhya.com Internet Source	<1%
5	Submitted to Università Carlo Cattaneo - LIUC Student Paper	<1%
6	Submitted to University of New South Wales Student Paper	<1%
7	Submitted to Regional Centre for Biotechnology Student Paper	<1%
8	Dipanjan Sarkar. "Text Analytics with Python", Springer Science and Business Media LLC, 2019 Publication	<1%



9	www.termpaperwarehouse.com Internet Source	<1%
10	arizona.openrepository.com Internet Source	<1%
11	es.scribd.com Internet Source	<1%
12	Submitted to Higher Education Commission Pakistan Student Paper	<1%
13	www.scribd.com Internet Source	<1%
14	www.inmybangalore.com Internet Source	<1%
15	theappsolutions.com Internet Source	<1%
16	student-friendly.blogspot.com Internet Source	<1%
17	Submitted to University of Southampton Student Paper	<1%
18	Submitted to essex Student Paper	<1%
19	Submitted to Visvesvaraya Technological University Student Paper	<1%



20	Submitted to Visvesvaraya Technological University, Belagavi Student Paper	<1%
21	erepository.uonbi.ac.ke Internet Source	<1%

Exclude quotes On

Exclude bibliography On

Exclude matches < 10 words

Publications in a Journal/Conference Presented/White Paper[2]

The publication of this work has been planned after the future work of deployment of state of the art API with dashboard.

Any Additional Details

GitHub link of Python Codes: https://github.com/sureshakukkaje/NLP---Voice-Of-Customers-Automotive-Gadgets

[2] URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.