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

Suresha K

BA05, SRN R19MBA11

Mentor : Akshay Kulkarni

Agenda

- Introduction
- Problem Statement
- Technical Solution Design
- Data Pipeline
- Framework – CRISP-DM
- Analysis
- Future Scope

Introduction

- Analytics – a megatrend today!
- Social networking sites – huge source for consumer voices
- Twitter and Facebook – ready to use information
- Data – can be analyzed using text analysis
- Can predict – sentiments towards the product
- Cost – reduced amount of cost
- Time – reduced turnaround time
- Big data – able to handle large amount of data

Problem Statement

“ 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 ”

Technical Solution Design

Data Extraction:

→Comments from Facebook

→Tweets from Twitter

Text Preprocessing / EDA:

→Stop Words

→Stemming, Lemmatization

→Standardize

→Text to Features

Modeling / Classifiers:

→Lexicon

→TextBlob

→VADER

→Support Vector Classifier

→Gradient Boosting

→Decision Tree

→Random Forest

→Gaussian NB Classifier

→Word Embedding – TF-IDF, Word2Vec

→Transfer Learning – Fine Tuning BERT

Validation:

→Accuracy

→Precision

→Recall

→F1-Score

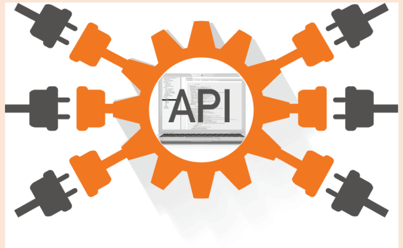
→Comparison of models

Phase - I

Deployment:

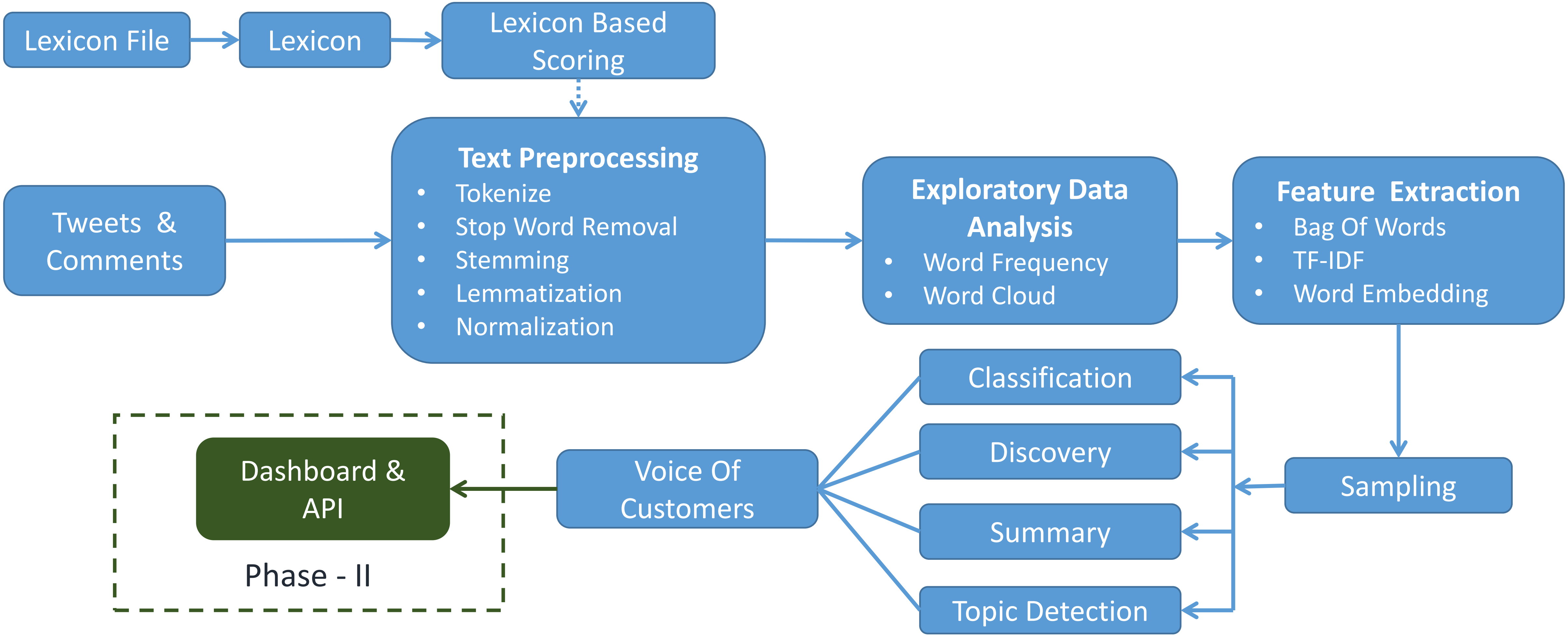
→Dashboard

→State of the art API



Phase - II

Data Pipeline



Framework – CRISP-DM

CRoss Industry **S**tandard **P**rocess for **D**ata **M**ining

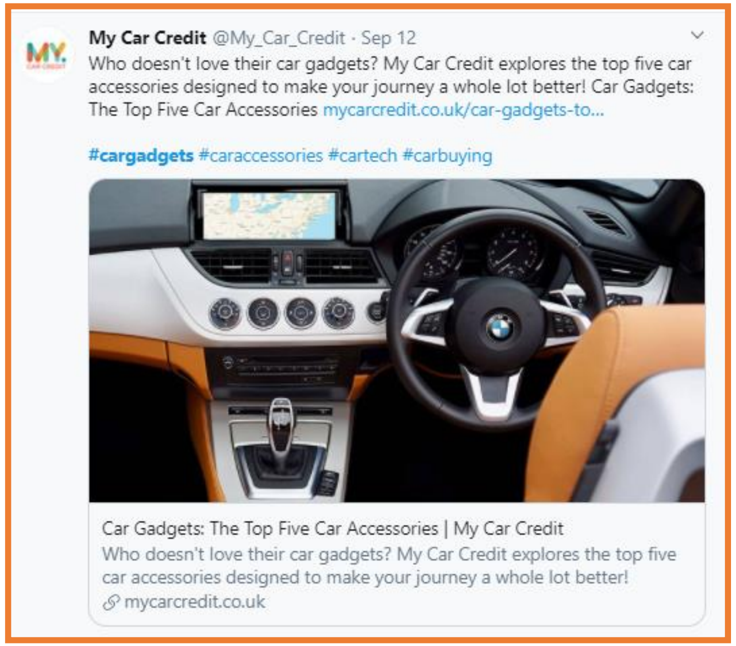
- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

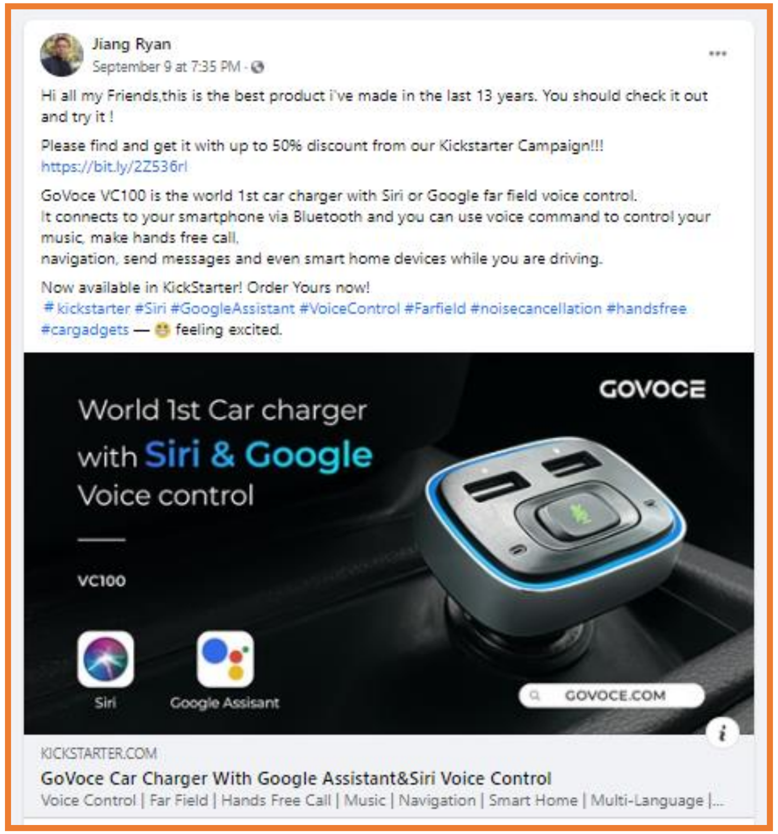
Business Understanding

- 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**
- 100 million users generating over **500 million tweets** every day
- **Facebook is used by a large** number of people on earth for expressing their ideas, thoughts, sorrows, pleasures and poems
- Facebook is the most widely adopted social media platform **by brands and companies**
- Extremely required for quick, agile decisions with business perspective **to use ML / AI technologies like NLP, Text Mining**
- It is imperative and **an opportunity** to use NLP as a front end API

Data Understanding

#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





- The user comments are normally attached with **hashtags** in any social network
- Hashtags help users to find the messages for **specific theme** or content

Data Understanding…

- Data extracted through **Twitter API** and **Export Comments APIs**
- The data has user comments for **9,546 rows** from Twitter and Facebook
- Comments/Tweets extracted for about **36 hashtags** related to user comments on automotive gadgets

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

Initial Comments Extracted

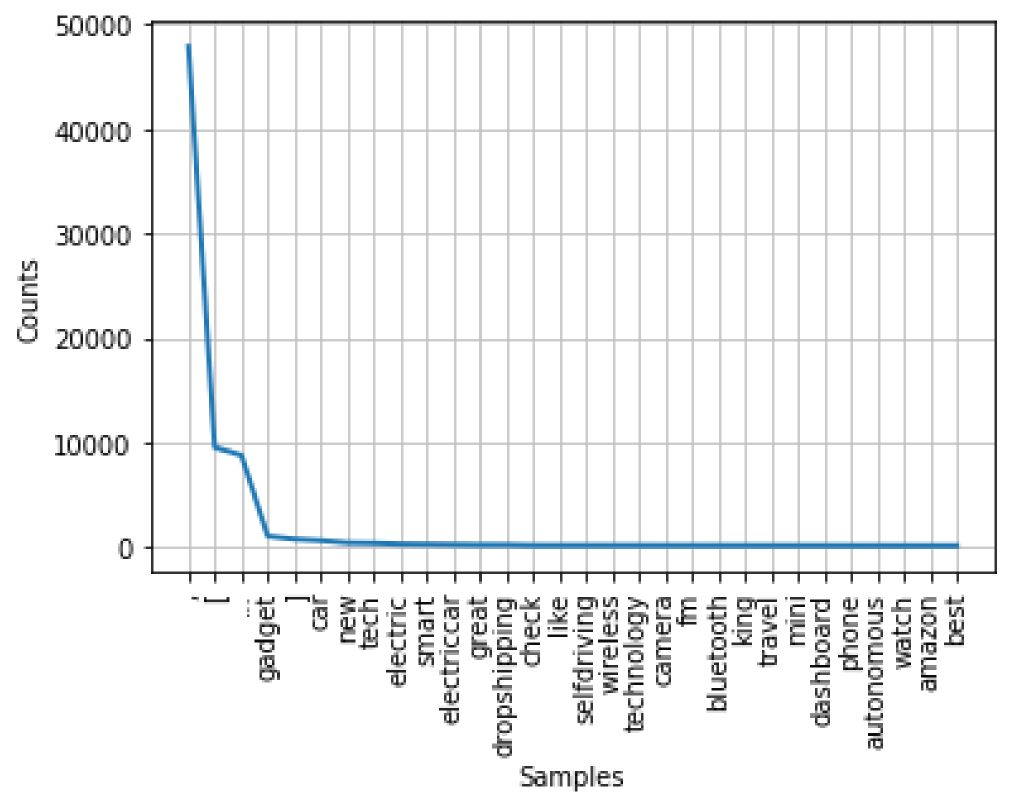
Comments	Word count	Char count	Avg word	Stopwords	Hashtags	Numerics	Upper
I just bought this projector and i am clueless...	16	77	4.133333	9	0	0	1
You promised me a multi card reader for leavin...	15	80	4.4	4	0	0	0
Hello Apeman, your team is not responding to m...	18	102	4.722222	8	0	0	0
Great picture, I am so excited about looking f...	38	245	5.777778	13	5	0	1
I have just bought the Apeman 550 dashcam and ...	32	154	3.967742	11	0	2	2

Understanding of Data

Data Preparation

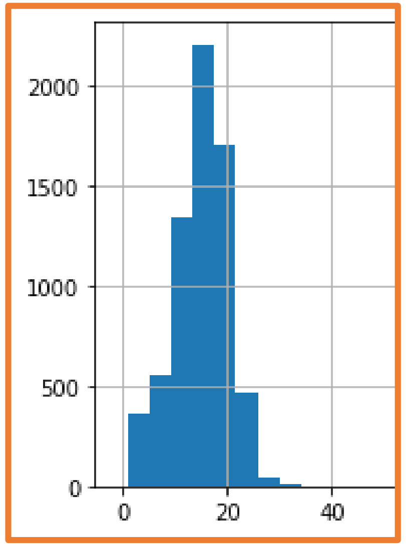
Text Preprocessing:

- Stop Words Removal
- Split sentences into words and lower case
- Handles Removal like @, URLs etc.
- Standardize: Full Forms, Lower Texts, Emojis
- Spell Corrections, removal of punctuations
- Join words to form the original sentences
- Tokenize
- Stemming and Lemmatization
- Create Corpus
- Bag of Words
- Word Cloud and Word Frequency



Frequency Distribution

Most Common Words	
gadget	2611
car	1031
tech	741
new	549
electriccar	509
technology	424
electric	432
autonomousvehicle	415
smart	378
watch	374
camera	372
travel	337
electron	336
selfdrive	283



Length of Comments

Data Preparation…

Text Preprocessing…

Comments	Cleaned messages	Tokenized messages
I just bought this projector and i am clueless...	I just bought this projector and i am clueless...	[i, just, bought, this, projector, and, i, am,...
You promised me a multi card reader for leavin...	You promised me a multi card reader for leavin...	[you, promised, me, a, multi, card, reader, fo...
Hello Apeman, your team is not responding to m...	Hello Apeman your team is not responding to my...	[hello, apeman, your, team, is, not, respondin...
Great picture, I am so excited about looking f...	Great picture I am so excited about looking fo...	[great, picture, i, am, so, excited, about, lo...
I have just bought the Apeman 550 dashcam and ...	I have just bought the Apeman 550 dashcam and ...	[i, have, just, bought, the, apeman, 550, dash...

Stop word removal	Stemmed messages	Lemmatized messages
[bought, projector, clueless, connect, iphone,]	[bought, projector, clueless, connect, iphon,]	[bought, projector, clueless, connect, iphone,]
[promised, multi, card, reader, leaving, revie...	[promis, multi, card, reader, leav, review, im...	[promised, multi, card, reader, leaving, revie...
[hello, apeman, team, responding, email, fixin...	[hello, apeman, team, respond, email, fix, iss...	[hello, apeman, team, responding, email, fixin...
[great, picture, excited, looking, 4k, cameraj...	[great, pictur, excit, look, 4k, camerajust, n...	[great, picture, excited, looking, 4k, cameraj...
[bought, apeman, 550, dashcam, find, rear, cam...	[bought, apeman, 550, dashcam, find, rear, cam...	[bought, apeman, 550, dashcam, find, rear, cam...

Word	TF	IDF
card	1	4.610001
reader	1	6.273506
promised	1	7.372118
still	1	5.450305
waiting	1	6.765982
multi	1	4.642089
im	1	2.355943
review	1	4.519487
leaving	1	8.065265
susie	1	9.163877

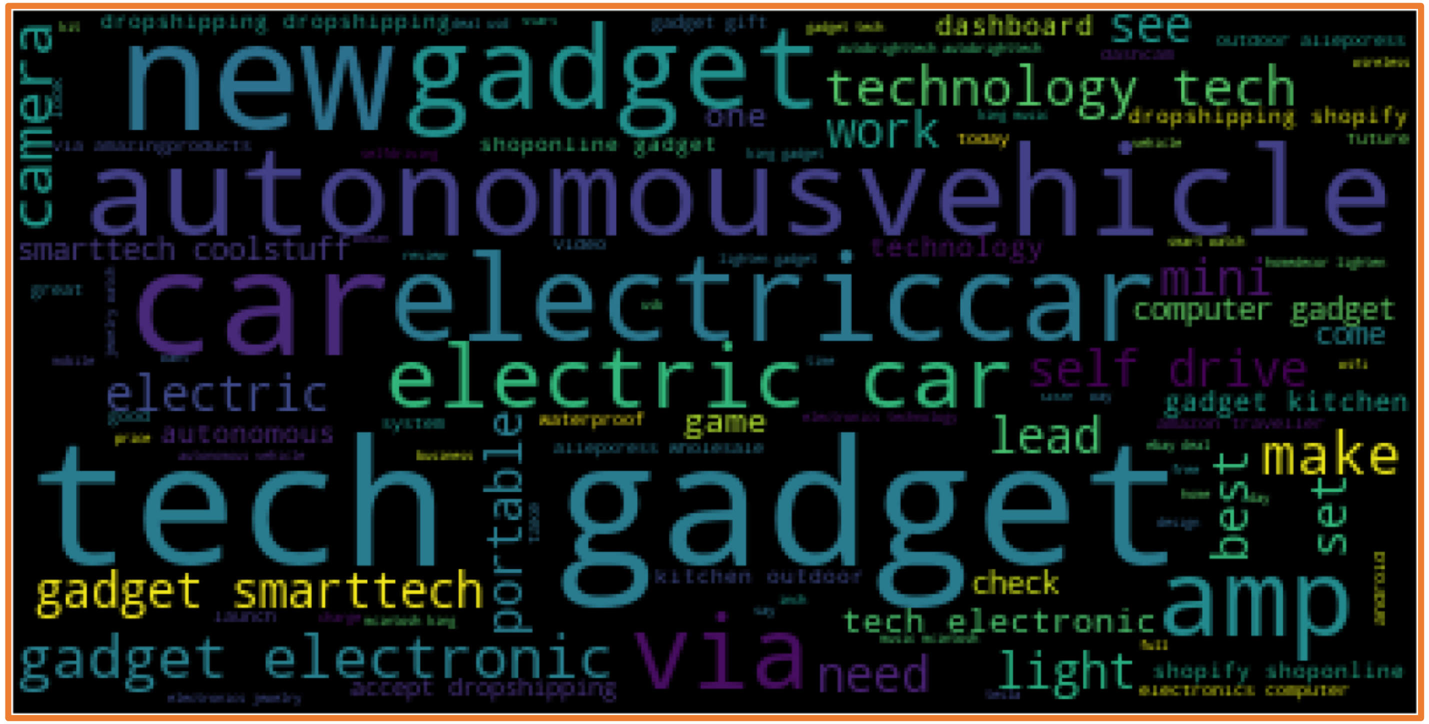
POS Tagging	
Full	NNP
link	NN
in	IN
bio	NN
dashcam	JJ
driving	VBG
fails	NNS
drivesafe	JJ
apemancamera	NN

	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]

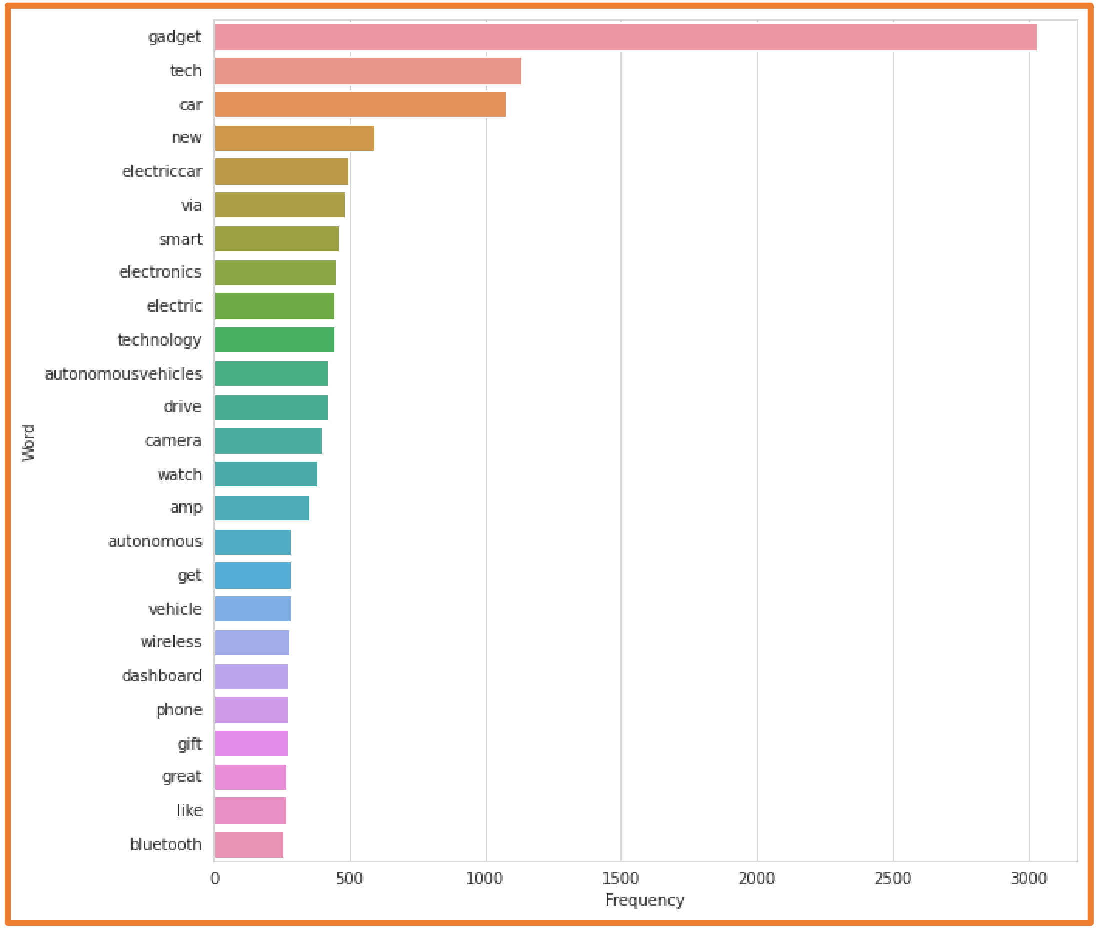
N-Grams	
bought	projector
projector	clueless
cluess	connect
connect	iphone

Data Preparation…

Exploratory Data Analysis:



Word Cloud

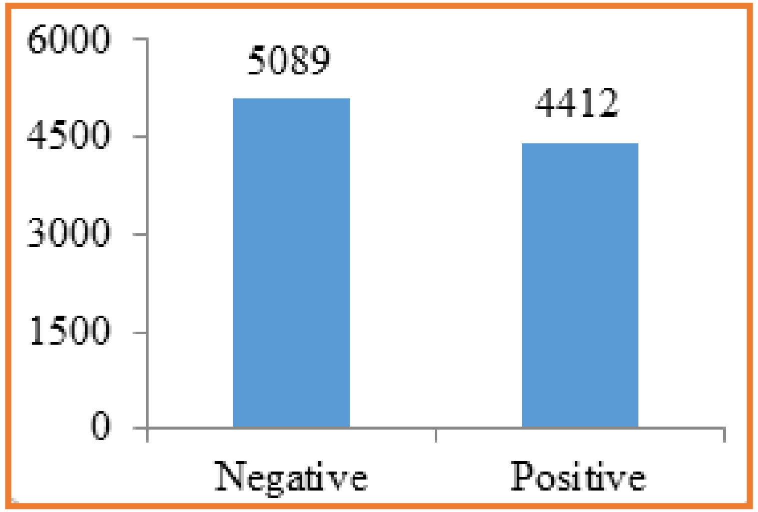


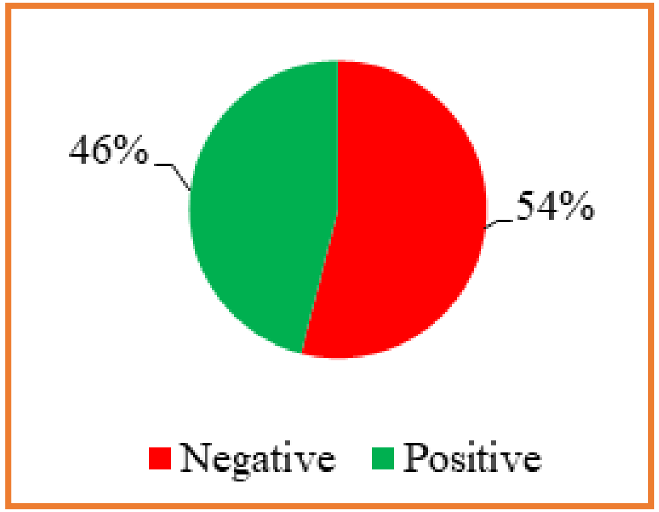
Word Frequency

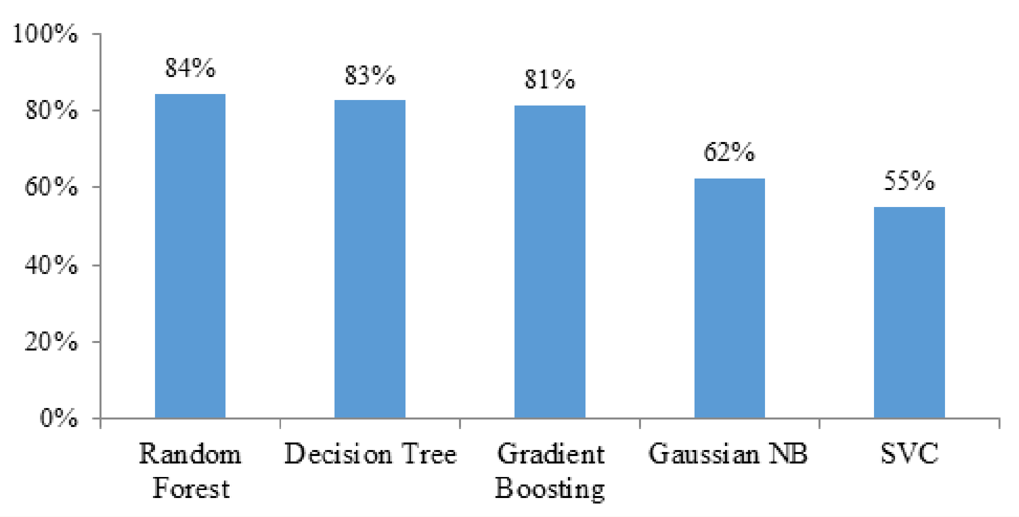
Modelling – Lexicon Approach

- Lexicon Model using **AFINN** (Finn Årup Nielsen)
- It has **over 3,300+ words** and it has a polarity score associated with each word

Comments	Score
bought projector clueless connect iphon	-2
great pictur excit look k camerajust need price come littl look forward f	3
bad news apeman put mine garbag done cheap product	-3
got one christma gift set work great day love littl research youtub figur	10
thank advanc jame hehe	2





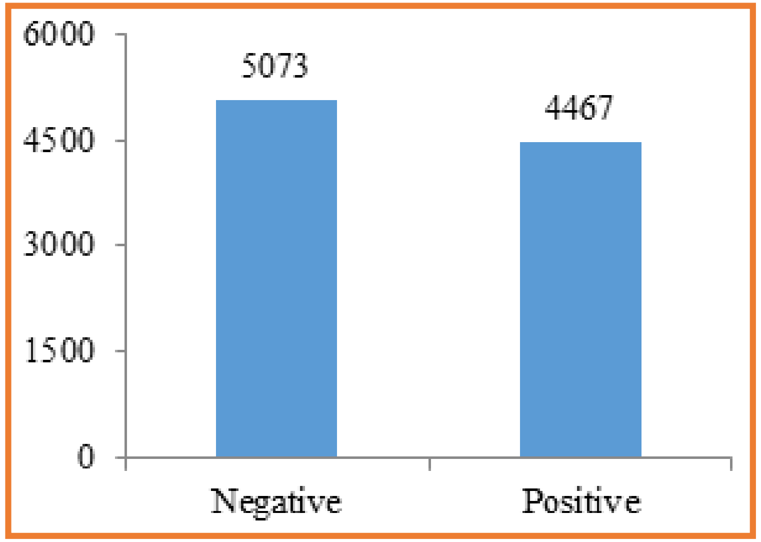


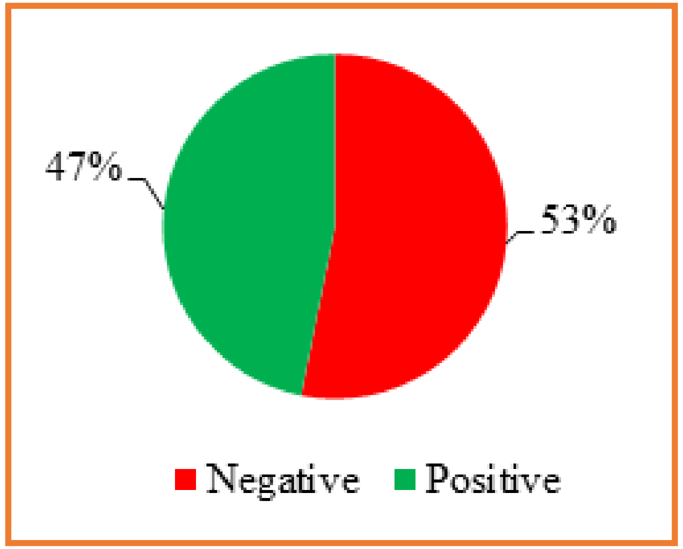
Modelling – TextBlob Sentiment

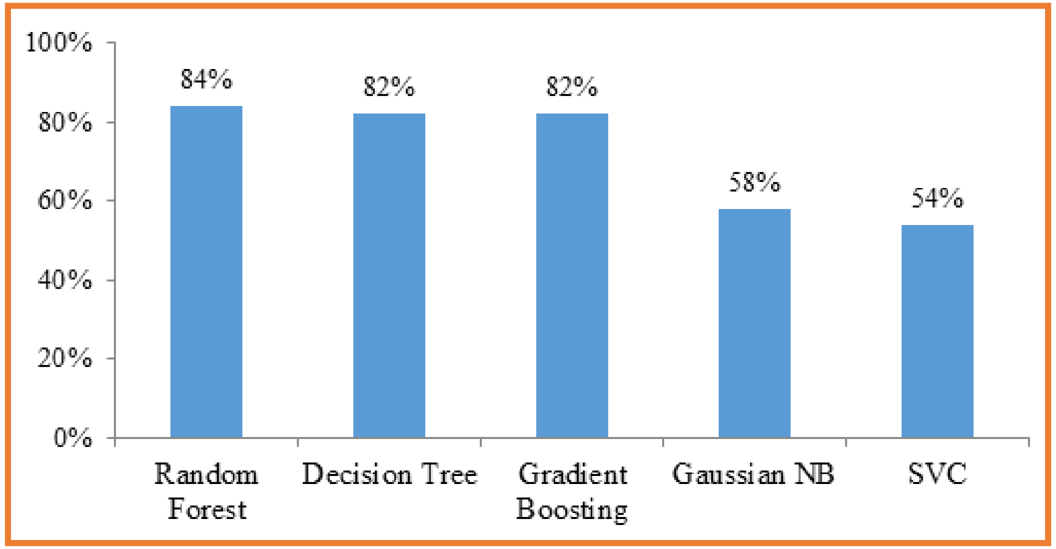
REVA UNIVERSITY

- Polarity using TextBlob
- The sentiment function of TextBlob returns two properties, **polarity and subjectivity**
- Polarity is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement

Comments	Polarity
bought projector clueless connect iphon	0
bad news apeman put mine garbag done cheap product	-0.15
great pictur excit look k camerajust need price come littl look forward	0.8
promis multi card reader leav review im still wait susi	0
bad news apeman put mine garbag done cheap product	-0.15
best new wifi mini dvr cam nanum vodool	0.568181818



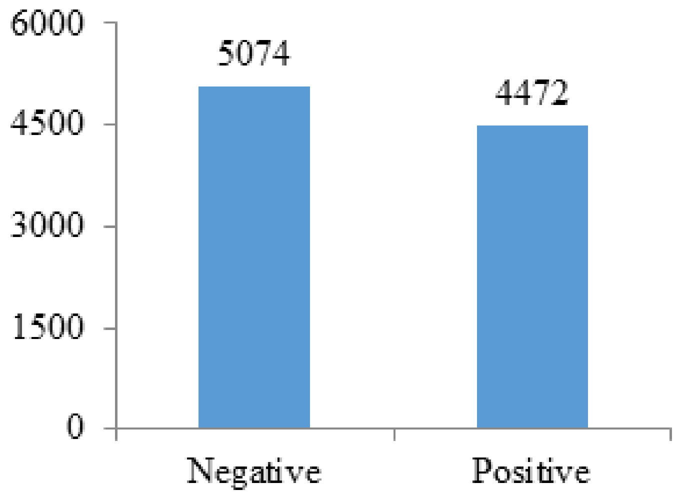


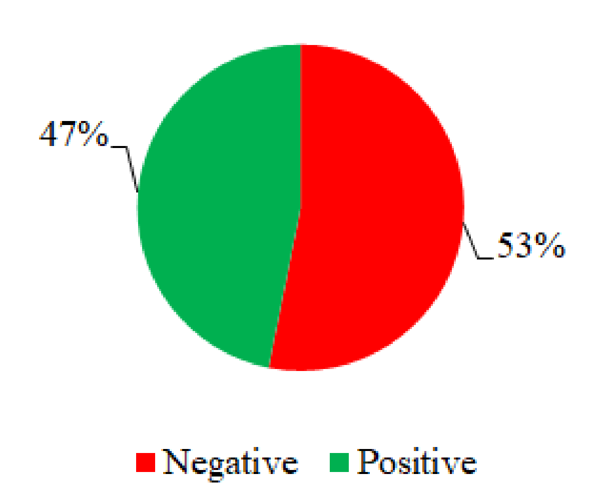


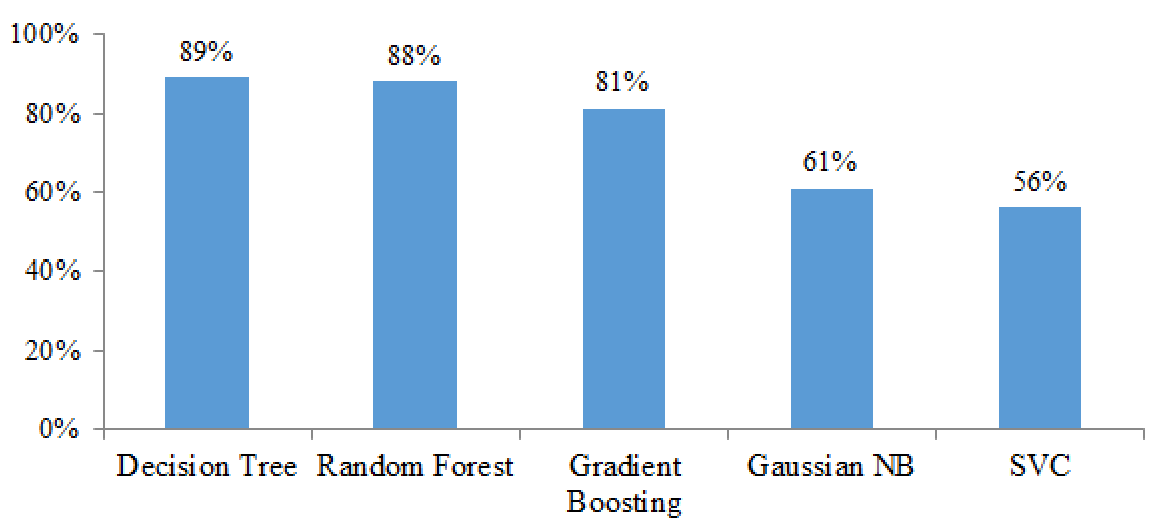
Modelling – VADER Sentiment

- Valence Aware Dictionary and sEntiment Reasoner
- It comes under **rule-based** systems
- Sensitive to the **polarity and intensity** of the emotions of the user
- Constructed from a **valence based** and able to generalize sentiment

Comments	Polarity
I just bought this projector and i am clueless how to connect t	-0.4329
You promised me a multi card reader for leaving a review I'm s	0.3612
Hello Apeman, your team is not responding to my email or fix	0
Great picture, I am so excited about looking for my own 4k car	0.863
Never give up. I can win this.	0.5859







Modelling – Word Embedding

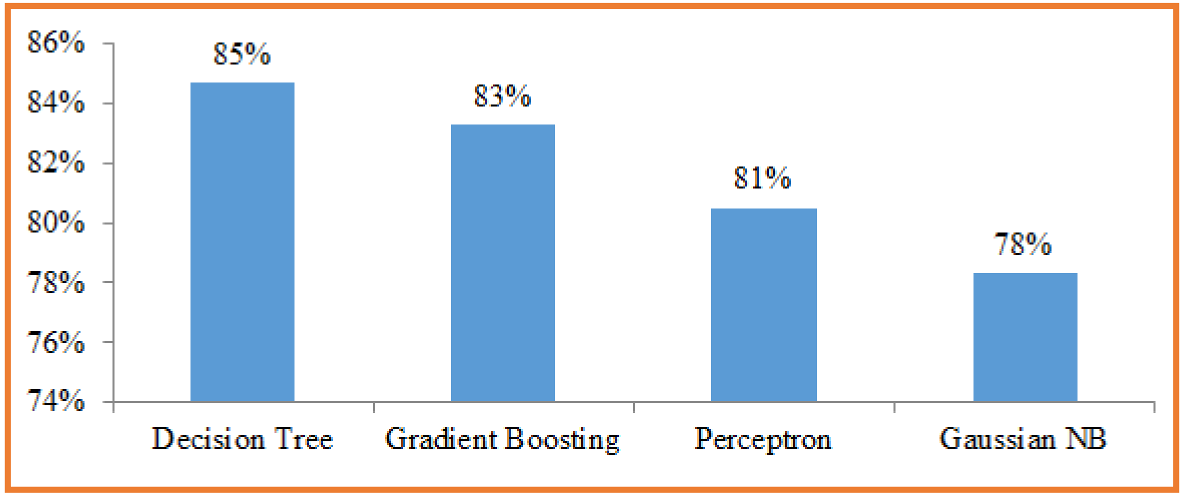
- New way of representing words into vectors
- Will redefine the high dimensional word features into low dimensional feature
- **Preserves the contextual similarity** in the corpus

TF-IDF Vectorizer:

- Converts the text documents into vector models
- Based on the **occurrence of words** in the documents without taking considering the exact ordering
- **Term Frequency** measures **how frequently a term occurs** in a document
- **inverse Document Frequency** measures **how importance the term is**

Word2Vec:

- Input for Word2Vec is a **text corpus**
- Produces the **word vectors** as output
- **Vocabulary** is constructed first from the training text data
- Learns vector representation of words
- Resulting **word vector will be used as features** in NLP applications



TF-IDF Vectorizer

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

Word2Vec

Modelling – Transfer Learning

- **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
- Introduced by **Google**

Fine Tuning with BERT:

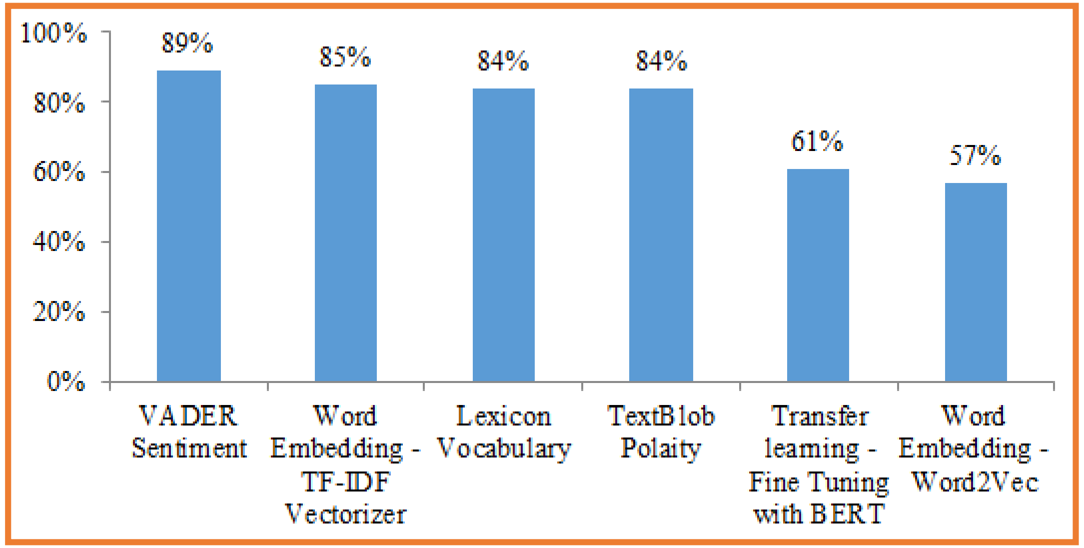
- Bidirectional Encoder Representations from Transformers
- First deeply bidirectional **unsupervised language representation**
- **Pre-trained** using plain **text corpus**

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

Evaluation

REVA UNIVERSITY

- **VADER sentiment** with **Decision Tree Classifier** has given the highest accuracy of **89%**
- **Word Embedding with TF-IDF** Vectorizer has the second best accuracy of **85%**
- The Precision, Recall and F1-Score of VADER Sentiment are as shown below:



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

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

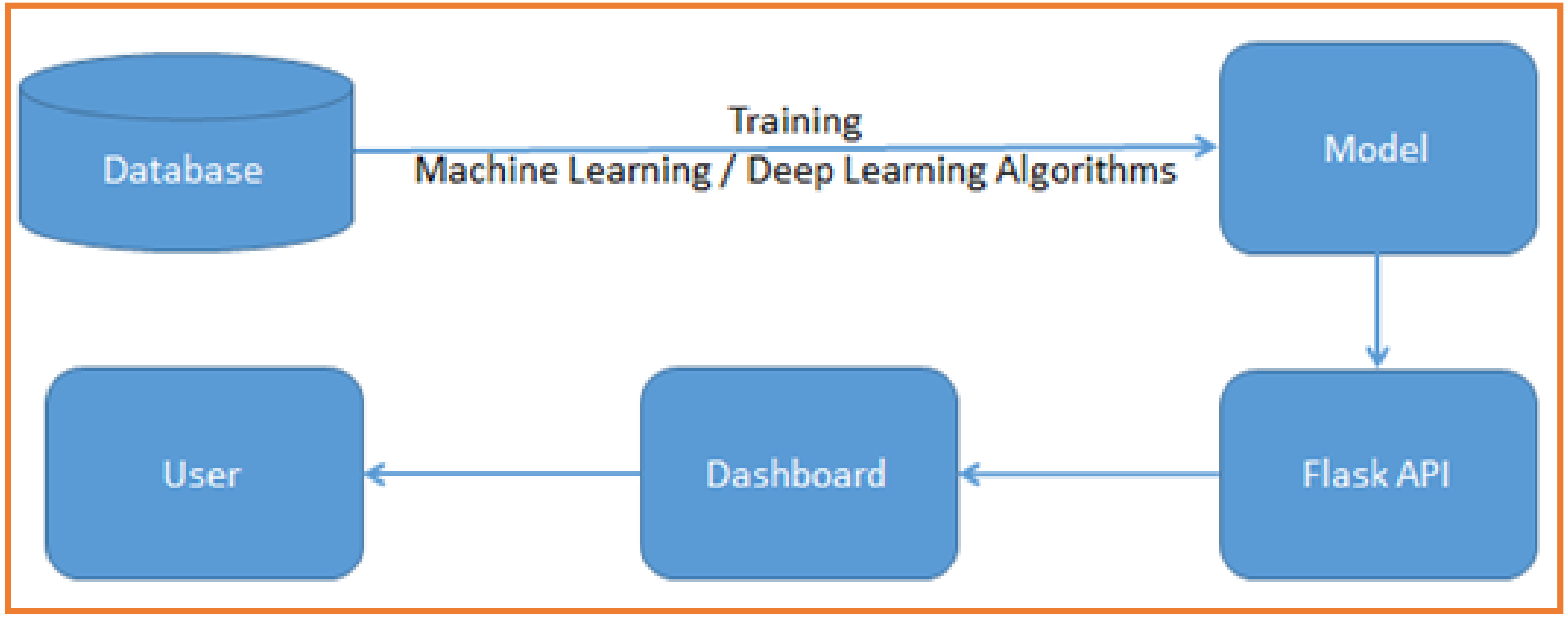
Results of VADER Sentiment

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

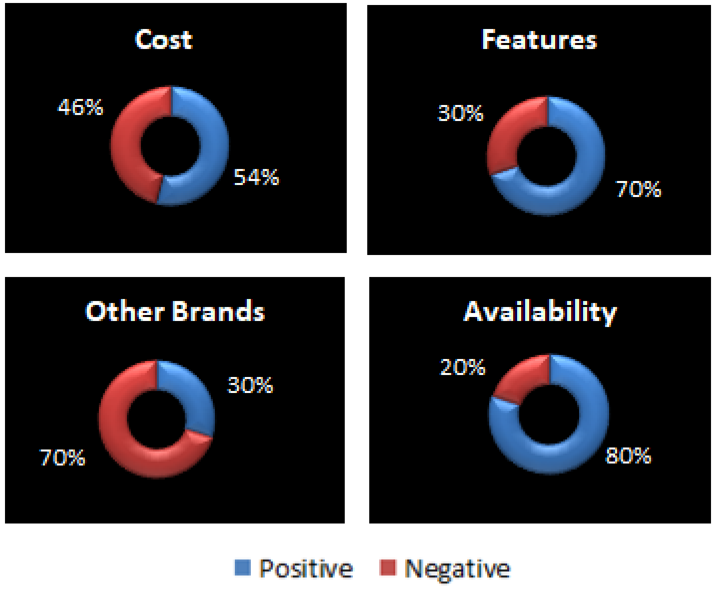
Comparison of Accuracies

Deployment

- As future plan in Phase – II, deployment of **state of the art API** and **dashboard** have been planned:



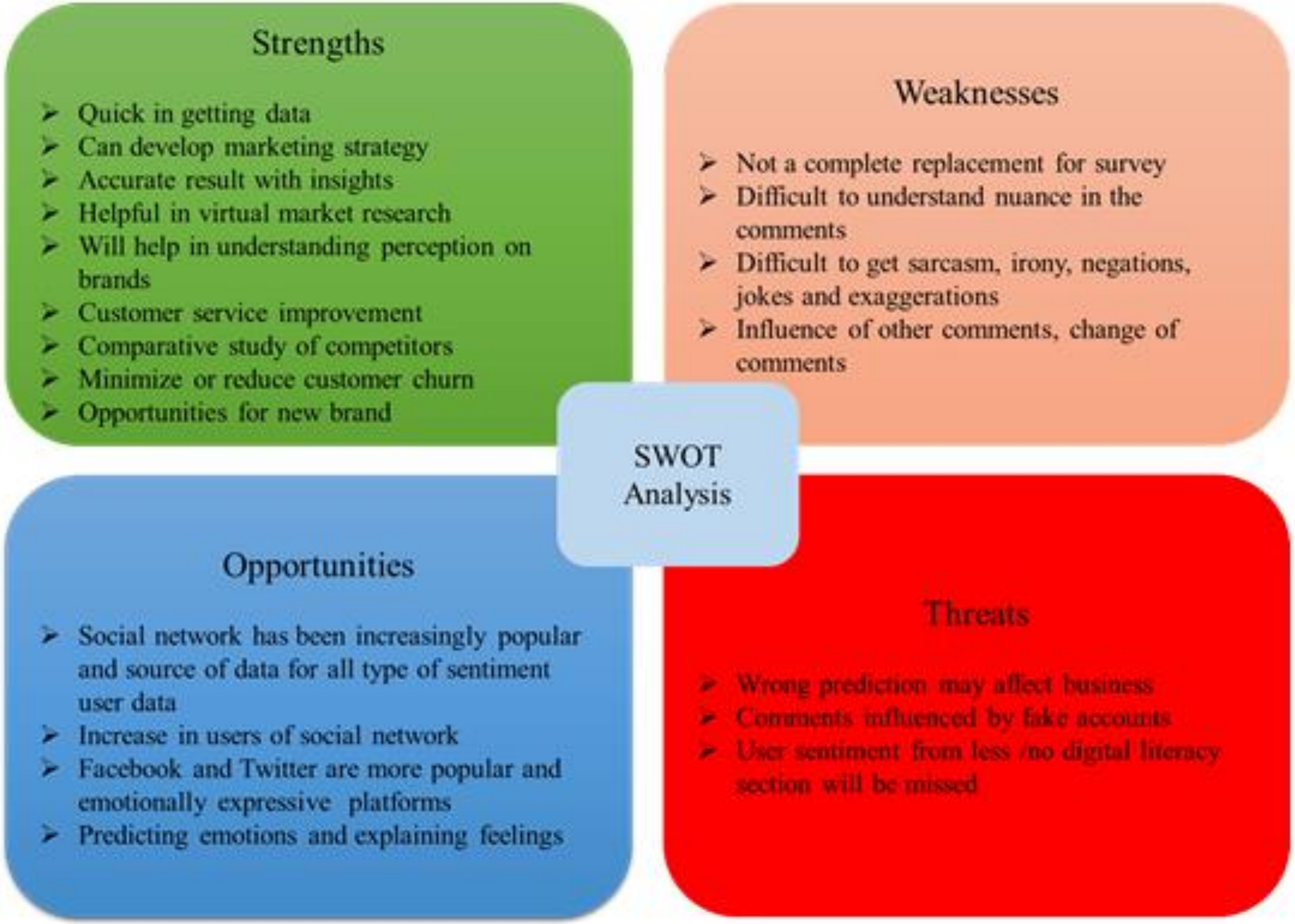
Deployment Plan



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

Dashboard Plan

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

Future Scope

- The model performance can be improved by focusing on:
  - 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
- **Emotion detection** and **predicting the comments**
- **Identify** comments from **fake accounts**
- **Dashboard** and **state of the art API**

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