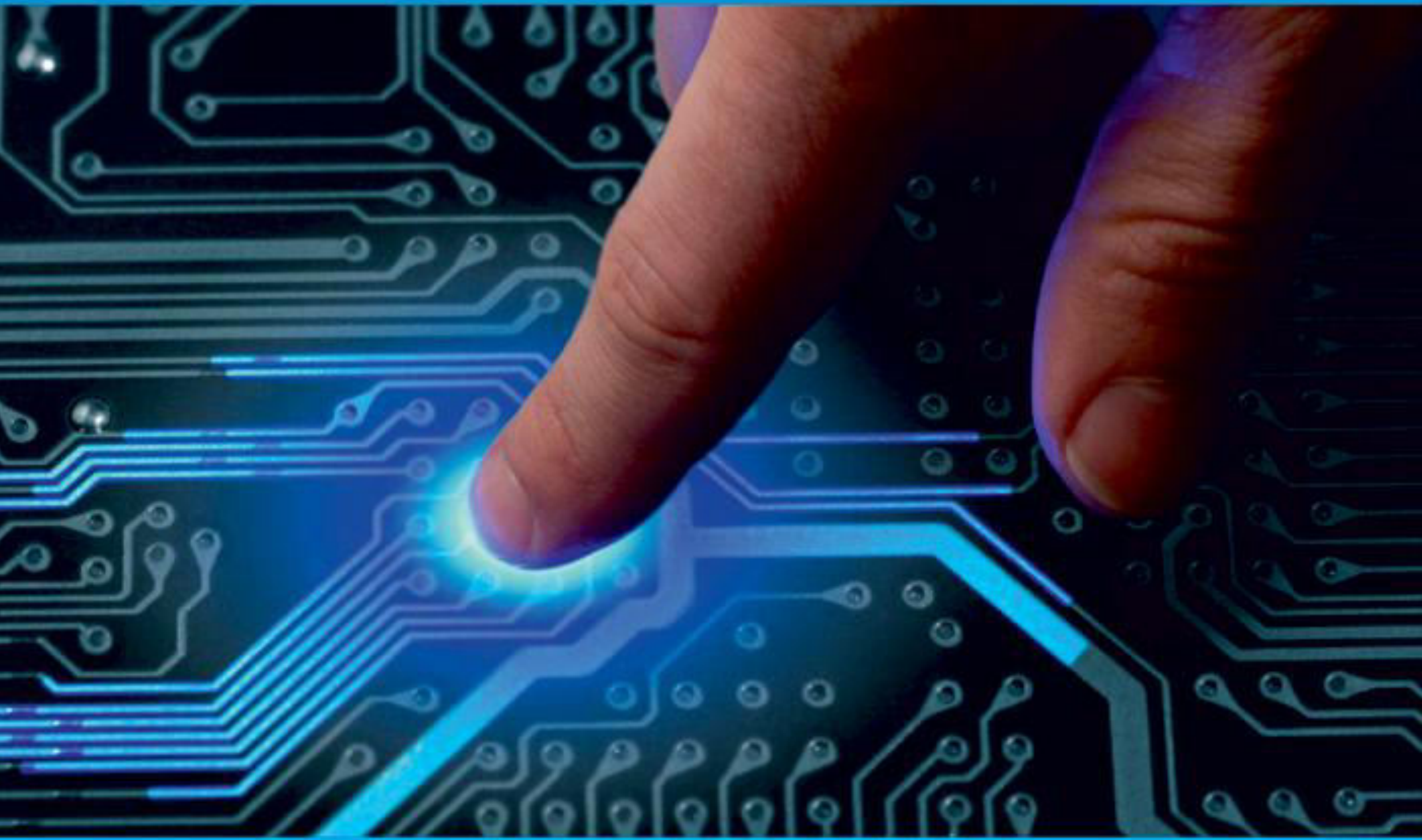
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

| DOI: 10.15680/IJIRCCE.2021.0906249 |

Topic Modeling on Luxury Car Reviews Using Natural Language Processing

B. Lakshmi Praveena[1], P.N.S.Mahesh[2], P.Jashwanth Kumar[3] , P.Madhu Kiran[4]

Assistant Professor, Dept. of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India[1]

B.Tech Student, Dept. of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India[2,3,4]

ABSTRACT: There has been a rapid growth in sales ofluxury cars over the last decade. So we wanted to understand what are the qualities that are most important to buyers through the customer reviews. This project is using Natural Language Processing and Latent Dirichlet Allocation for topic modeling and sentiment analysis of the reviews of 5 luxury car brands.

On the customer reviews the Natural Language Processing and Latent Dirichlet Allocation are applied to understand the important qualities ofcars that are helpful for buyers while purchasing cars. This paper also predicts the sentiment of the reviews.

KEYWORDS: Latent Dirichlet Allocation, NaturalLanguage Processing, Sentiment analysis,Topic modeling.

I. INTRODUCTION

Topic Modeling on luxury car reviews is a dataanalytics application which can be used to know the qualities that are most important to buyers through the customer reviews. This application extracts meaningful topics from the reviews and theresult computed.

To use this application user has to upload the dataset consist of car reviews. This application performs sentiment analysis on the extracted topics from reviews. It represents the extracted topics in HTML format and also calculates the sentiment polarity and compares the sentiment polarity of different cars using a graph.

II. PROPOSED METHODOLOGY OF TOPIC MODELING

In the proposed system of Topic Modeling on luxury car reviews, with the help of the customer reviews it provides the best topics with high coherence values. It uses both LDAMallet and LDAMulticore for topic modeling. So, on the bases of the runtime and the coherence values it extracts the topics.This application provides the best topics on comparing the results of LDAMallet and LDAMulticore outputs. There are mainly three modules in this application. The modules of this application are as follows:

1. Preprocessing module:

In this module the input dataset is preprocessed with preprocessing methods and the process of preprocessing includes following steps.

- Removing null values or missing values. Deletes the rows with no reviews.
- Dropping unwanted columns or attributes which are not related to the topic modeling processand sentiment analysis process.
- Changing the data type of the date column.
- Removing stop-words with NLTK library.
- Removing numbers from text with regular expression function.
- Change text to Lower case and remove words less than 3 letters.

2. **Topic modeling module:**

At this module the visualization process starts. e.g. word cloud to understand the common wordsin the review. It creates a dictionary and a corpus with the review text which are needed for the LDA models. LDA Multi-core and LDA Mallet are used to extract the meaningful topics from the text.

3. Sentiment Analysis & Visualization module

It is the module which conducts sentiment analysis on the extracted topics and displays the output. The output consists of a bar graph in which sentiment polarity is plotted and the .CSV file which has the percentage ofpositive reviews, negative reviews and neutral reviews.

III. TOPIC MODELING

In this paper, we propose a topic modeling scheme that uses both LDAmallet and LDAMulticore Mallet for topic modeling to see the one with the highest coherence score and with meaningful topics from the reviews [1].

In order to extract the meaningful topics from the customer reviews we use topic modeling andsentiment analysis [2].

1. Topic Modeling:

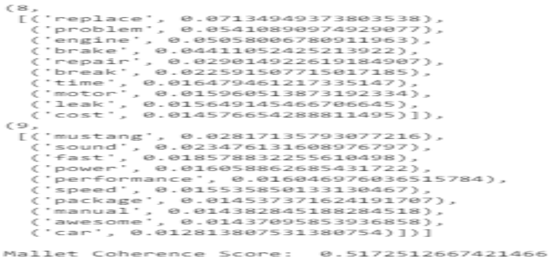
The topic modeling analysis refers to a series of techniques that identify what text deals with. Topic modeling is a technique that extracts and suggests potentially meaningful topics from a great number of documents based on a procedural probability distribution model.

We use the LDA models for topic modeling. Latent Dirichlet allocation (LDA) is the most popular topic model, which is a method for analyzing alarge set of documents [4]. The basic idea is that documents are represented as a topic distribution where each topic is characterized by a word distribution. This application uses both LDAMallet and LDAMulticore.

2. LDAMallet:

Mallet, an open source toolkit, was written by Andrew McCullum. It is basically a Java based package which is used for NLP, document classification, clustering, topic modeling, and many other machine learning applications to text. It provides us the Mallet Topic Modeling toolkit which contains efficient, sampling-basedimplementations of LDA as well as HierarchicalLDA. It is fast when compared to LDAMulticore.

The MALLET topic model package includes an extremely fast and highly scalable implementationof Gibbs sampling, efficient methods for document-topic hyper parameter optimization, and tools for inferring topics for new documents given trained models [3].



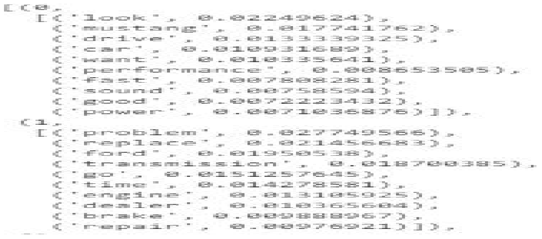
```
(8,
 [('replace', 0.07134949373803538),
  ('problem', 0.05410890974929077),
  ('engine', 0.05058006780911963),
  ('brake', 0.04411052425213922),
  ('repair', 0.029014922619184907),
  ('break', 0.022591507715017185),
  ('time', 0.016479461217335147),
  ('motor', 0.015960513873192334),
  ('leak', 0.015649145466706645),
  ('cost', 0.014576654288811495)]),
(9,
 [('mustang', 0.02817135793077216),
  ('sound', 0.023476131608976797),
  ('fast', 0.018578832255610498),
  ('power', 0.016058862685431722),
  ('performance', 0.016046976036515784),
  ('speed', 0.015535850133130467),
  ('package', 0.014537371624191707),
  ('manual', 0.014382845188284518),
  ('awesome', 0.01437095853936858),
  ('car', 0.012813807531380754)])]

Mallet Coherence Score:  0.5172512667421466
```

Fig:- Output format of LDAMallet

3. LDAMulticore:

It is also an open source toolkit, which is used for NLP, document classification, clustering, topic modeling, andmany other machine learning applications to text. It uses all CPU cores to parallelize and speed up model training [5].

The LDA module in gensim is very scalable, robust, well tested by its users and optimized in terms of performance, but it still runs only in single process, without full usage of all the cores of modern CPUs.



```
[(0,
  [('look', 0.02249624),
   ('mustang', 0.017741762),
   ('drive', 0.013339325),
   ('car', 0.010931689),
   ('want', 0.010335641),
   ('performance', 0.008653505),
   ('fast', 0.007808281),
   ('sound', 0.00758594),
   ('good', 0.007222343),
   ('power', 0.007103687)]),
 (1,
  [('problem', 0.027749566),
   ('replace', 0.021456683),
   ('ford', 0.01950538),
   ('transmission', 0.018700385),
   ('go', 0.015125764),
   ('time', 0.01427858),
   ('engine', 0.013105925),
   ('dealer', 0.010365604),
   ('brake', 0.009888967),
   ('repair', 0.009769211)]),
```

Fig :- Output format of LDAMulticore

With the help of word cloud we are visualizing top words, and keywords that are present in the reviews. Word Cloud analysis shows that excluding'car, 'truck, 'mile, 'ford and, 'vehicle' are more frequently mentioned in online reviews of 5 car brands named 'genesis', 'dodge', 'Ferrari', 'fiat', 'ford'.

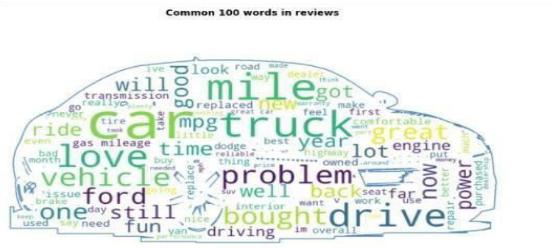


Fig :- Common 100 words in reviews

Looking for the key topics shown in the entire document of the customer's reviews was classified by subject, and judged that the most appropriate topics were classified into eight[6].

Among the words derived from each graph of eight topics, the high beta value has the most important meaning in that topic. For example Each topic is a representative theme of 10 keywords and Topic 1 isnamed'Look of cars' which can be explained by words such as 'look, 'mustang', 'drive' etc. Topic
2 was named 'Car airbag and safety' with keywords like 'problem', 'replace', 'time', 'engine' etc. which can represent the inside of the car.
The result of the topic modeling is represented in HTML format. The html file is generated by using the module **pyLDAvis.**

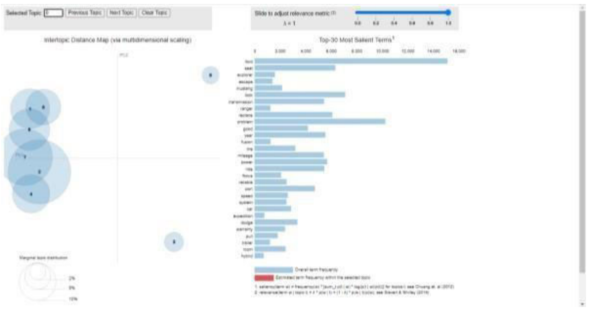
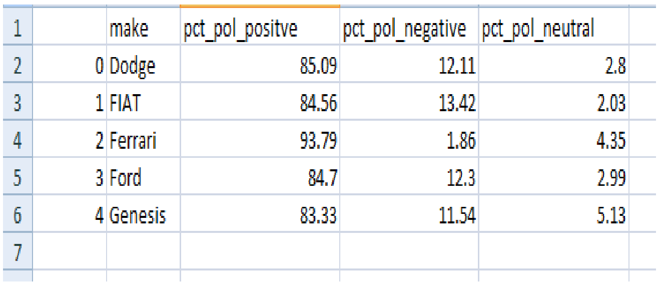

Selected Topic
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric
Intertopic Distance Map (via multidimensional scaling)
Top-30 Most Salient Terms
Overall term frequency
Estimated term frequency within the selected topic
Marginal topic distribution

Fig :- Output of Topic Modeling

IV. SENTIMENT ANALYSIS

Sentiment analysis refers to a technique that classifies or quantifies emotions in text and turns them into objective information. Humans use language to communicate their thoughts and feelings. If the topic modeling introduced earlierwas a text mining technique that identifies the "target covered by text", the sentiment analysis is a text mining technique that estimates the “attitude contained in text”. Just as the topic modeling extracts words that embody topics assumed to be inherent in the text and estimates topics, sentiment analysis also estimates the feelings inherent in the text.

In this paper we are generating the percentages of positive poll, negative poll and neutral poll present in the reviews. We are also generating the sentiment polarity distribution graph in which sentiment frequency is plotted with polarity.



1		make	pct_pol_positve	pct_pol_negative	pct_pol_neutral
2	0	Dodge	85.09	12.11	2.8
3	1	FIAT	84.56	13.42	2.03
4	2	Ferrari	93.79	1.86	4.35
5	3	Ford	84.7	12.3	2.99
6	4	Genesis	83.33	11.54	5.13
7					

Fig :- CSV file of Sentiment Analysis.

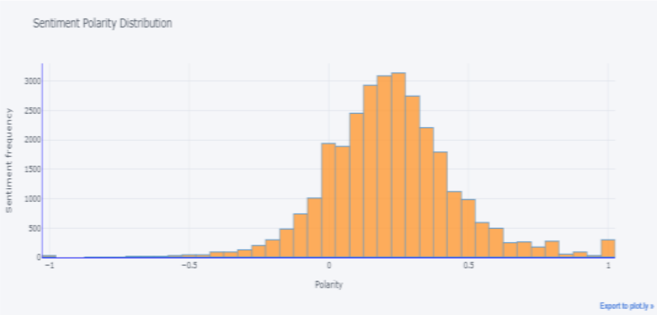


Fig :- Graph of sentiment polarity Distribution

V. CONCLUSION

The topic modeling on the luxury car reviews is developed to facilitate the manufactures with the information of what customers saying about cars from their reviews. It also provides the information of in which perspective the users are satisfied and in which perspective they are not satisfied. So, with the help of the topics extracted and the polarity manufactures try to increase the quality and also try to satisfy the users and increases their sales.

VI. FUTURE SCOPE

This project Topic Modeling on luxury car reviews has been developed in such a manner, that the futurerequirements of the company are met. The project is flexible to adapt the changes efficiently without affecting the present system. In future, there can be a provision of reading the reviews automatically from the webpage. And also checking polarity based on the context and also depicting the sarcastism in the reviews.

We are also planning to implement the web application which takes car company as input which directly reads the reviews and displays the results. This is the future scope of our project.

REFERENCES

1. https://www.tutorialspoint.com/gensim/gensim_topic_modeling.htm
2. https://www.kaggle.com/datasets/
3. http://mallet.cs.umass.edu/
4. https://towardsdatascience.com/to pic-modeling-and-latent- dirichlet-allocation-in-python- 9bf156893c24
5. https://radimrehurek.com/gensim/models
6. /ldamulticore.html
7. https://researchleap.com/consumers-brand-choice-behavior-for-luxury- cars-in-china

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details