Car Recommendation System Using Customer Reviews

Saumya Singh[1], Soumyadeepta Das[2], Ananya Sajwan[3], Ishanika Singh[4], Ashish Alok[5]

[1,2,3,4,5]*Vellore Institute of Technology, Tamil Nadu, India*

---***---

Abstract - *Consumers face a major challenge today in choosing from the many alternatives available to any product category. With the growing demand and production of cars, there are thousands of great brands which make hundreds of new models every year. The need for proper recommendation of the cars based on customer's specific needs is essential and acts integral to both the manufacturers as well as clients. Recommendation based on specifications and details, are becoming somewhat unrealistic and not everyone has the detailed knowledge of cars. Clients prefer to have a review-based recommendation from the people who have hands on experience with the particular product. We can see how recommendation systems have a surprisingly large impact on the materials consumers engage with over the course of their daily lives. Hence, our proposed solution is to provide a system for car recommendation based on customer reviews, using the power of ML, NLP and Data Analytics.*

Key Words: **Machine Learning, Natural Language Processing, Data Analytics, Data Visualization, Recommendation, Recommendation System**

1. INTRODUCTION

As the world's population grows, so does the value of the product in the market. Due to global distribution, there is an increase in global trade leading to different types of products e.g., buy soap, there are different types depending on the flavours, aroma, brand (international) etc. it also applies to cars.

As the number of cars on the international market increases, so does the information that each individual gains from an online product. People's madness with cars dates back to the Neolithic period, the last part of the Stone Age when the making of a wheel was made. As time goes by, technology grows and stands on what we see today. Today most people are aware of what is happening around them. As market competition grows, cars with similar features enter the market. People will be confused about what to choose. Here the recommendation algorithm plays a role because it assists the customer or end user in promoting the right product based on its taste.

This research project is about a web-based program for vehicles. The existing systems that rely on recommendation based on specifications and details, are becoming somewhat unrealistic and not everyone has the detailed knowledge of cars. Different customers use different strategies, some are more knowledgeable, and up to date whereas some need advice, reviews from peers, and advice. Our goal is to bring the best out of both worlds, be it specific search, or recommendation, be it seller claims, or reviews. Existing solutions require the users to have proper knowledge of specification and their requirement and, mostly are based on the content type recommendation. Using the customer reviews, our system becomes more customer oriented as well as less knowledge driven but is not confined to that.

The main purpose of this function is to recommend the car according to the user model and object profile. In this paper, a proposed algorithm to recommend hybrid-based user-to-user and interactive filtering techniques is used based on actual textual data from the end users or clients of the product, using the ML, NLP and Data Analytics.

The dataset for the project is **Consumer Car Reviews dataset from Edmunds.com (also available on Kaggle.com)**, as well as other sources, which contains lakhs of reviews from multiple brands given by consumers.

2. LITERATURE REVIEW

1. T. G. Thomas 1, V. Vaidehi , worked on *"Vehicle Recommendation System Design The Web uses a Hybrid Recommender Algorithm"* developed a web based complimentary program for vehicles. The main purpose of this function is to recommend the car according to the user model and object profile. In this paper, the proposed hybrid recommendation algorithm is used from user-to-user and interactive filtering methods aimed at generating vehicle recommendations. The user model is built with personality features, click data and browsing history. The profile of the item is built using various car attributes. Forty car types are used including 224 car types in this project.

2. Srivastava, A. Kumar, S. Samee, P. Thokal Vijay, P. S. Tanesh, worked on *"Vehicle Recommendation Method Using Vengatesan K Machine Learning Algorithm"* studied and found that more than 90% of planned drivers regularly show that eliminating any driving pollution can control their chances and adaptability. Competitive drivers have exerted pressure on the low-

key concept of open-vehicle integration. This weight conveys the idea that it is actually improving in terms of how half of the late respondents had eliminated any driving distractions that they felt were an open car that, or in some way, lacked.

3. G. Prabowo, Md. Nasrun, R. A. Nugrahaeni worked on *"Recommendations for the Combined Filtering Program (CF)",* proposed a program that can help provide information about vehicles that are in line with user preferences, i.e., a recommendation system. The recommendation system requires appropriate recommendations In this study you will focus on the problem of recommending a car selection system by creating a recommendation system using a collaborative filtering process.

3. DRAWBACKS AND CHALLENGES

Recommendation systems are an important part of business and e-commerce including the auto industry. The currently available car recommender systems depend heavily upon users having proper knowledge of cars and their specifications.

Most of the systems use a content-based recommender approach which is based on the user's history and suggests items similar to their past purchases. This type of system is extremely disadvantageous as users could be first-time buyers and may not have proper knowledge on car specifications.

These drawbacks have been eliminated with our application which uses a collaborative filtering approach i.e., "people to people" approach. Collaborative filtering methodology is applied to filter products that might interest a particular user depending on reactions received by similar users. This method is more suited to recommend cars as it is less knowledge-driven and more customer-oriented, keeping in mind that not all customers are informed on specifics of cars.

Challenges: One of the challenges faced was cleaning the user reviews available in the datasets. This was done by incorporating different methods of natural language processing like lemmatization, removal of stop words and punctuations. The datasets were combined onto a single data frame and cleaned by additionally removing unnecessary columns.

Another challenge included integrating the machine learning notebook with the Flask application to create a fully functional product. The challenge here was to ensure that the recommendation was fast, accurate and efficient. This involved multiple testing scenarios and removing bugs which improved the accuracy and the efficiency of the recommender system.

4. REQUIREMENTS & PROPOSED SYSTEM

4.1 Software Requirements

- Python (as programming language version>=3.0)
- Python compiler
- Conda or Jupyter environment (used for development)
- Flask
- Web Browser

4.1.1 Python Module Dependencies

- nltk
- genism
- PyLDAVis
- Sklearn
- Pandas
- Numpy
- Pickle
- Json
- re
- Random
- Textblob
- Math

4.2 Proposed System:

The Recommendation system for the car proposed involves development processes like Data Collection, Data Pre-processing, Model Design, Model Building, Recommendation. The system architecture involves Python as the main programming language, for the recommendation model.

The architecture involves the dataset collected from Edmunds Car Reviews dataset, which consists of the details of the reviewer, the car reviewed, the text review and the rating given by the customer for the car. The dataset is pre-processed which involves processes like stop word removal, lemmatization, removal of common words, etc. After pre-processing, the processed dataset is fed to Topic Modelling model using Non-Negative Matrix factorization after, count vectorization and inverse document factorization, after this stage the topics are extracted and distributed based on the dataset and the reviews on each car. To optimal number of topics in the dataset to avoid overfitting and high variance decisions, we check the coherence values vs number of topics, the number with the highest coherence value is taken as the number of topics. Upon deciding the number of optimal topics, the topics and the output is fed to the Latent Dirichlet Allocation (LDA) model, to score each record and find out the underlying relationship of a particular query with the review of the car, as a result of which a specific topic number is be allocated to each car, which is used for recommendation based on whatever query the user provides.

The application allows both quantitative searches based on car type like SUV, sedan and also qualitative search like how a user describes their requirements to an expert, who in return uses their expertise to recommend their cars. This application and recommendation model provides the capability for both.

Stages:

1. Data Collection

2. Data Pre-processing

3. Model Building:

    a. Topic Modelling using NMF

    b. Selecting optimal number of topics

    c. LDA model training

    d. Assign topic scores to cars and mapping back to dataset

4. Recommendation

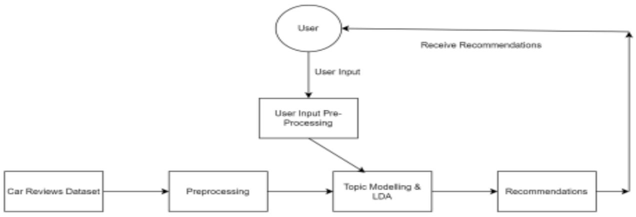


Fig -1: Architecture Diagram

5. MODULES DESCRIPTION

a) Dataset: The dataset used for the project is Consumer Car Reviews dataset from Edmunds.com (also available on Kaggle.com), which contains lakhs of reviews from multiple brands given by consumers. Contains datasets pertaining to different companies, which could be used separately for each company or combined to be used as whole general database. The usability of the dataset is marked 7.1-7.5, which is decent, and means it has low null and ill formatted data. Hence, as the source provides a huge data, along with proper columns and fields and has high usability score, we decided to work on this dataset. Consumer Car Reviews Dataset contains a huge database of car reviews from varying brands.

Contains 7 columns namely:

- Key,

- Review Date,

- Author Name,

- Vehicle Title,

- Review Title,

- Review, Rating.

The models will be trained based on the Date, Vehicle Name, Review Title, mainly the Review and Rating.

b) Data Pre-processing: The datasets are combined and loaded onto a single data frame and then we only consider data after 2010 due to relevancy. The dataset is cleaned by removing unnecessary columns and then we are left with the main data – Vehicle, Review and Rating columns.

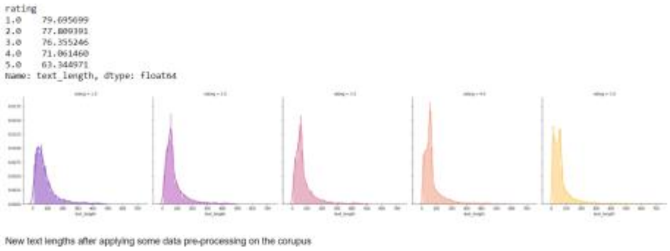
Data Pre-processing techniques like:

- Stop Word Removal

- Lemmatization

- Removal of common basic words

One important inference during pre-processing stage was, cars with higher rating had reviews with lesser text review length. Even after cleaning and pre-processing the dataset, it was holding true, and is used for model building.



```
rating
1.0    79.695699
2.0    77.809391
3.0    76.355246
4.0    71.061460
5.0    63.344971
Name: text_length, dtype: float64
```

New text lengths after applying some data pre-processing on the corupus

Fig -2: Preprocessing Inference (ratings inversely proportional to review length)

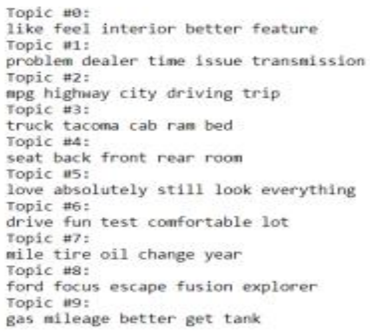
c) **Topic Modelling:** The processed data is used for training the topic modelling model, topic modelling is done using the NonNegative Matrix Factorization (NMF) model from sklearn, which takes the input matrix and outputs two matrices

• The W factor contains the document membership weights relative to each of the k topics. Each row corresponds to a single document, and each column correspond to a topic.

• The H factor contains the term weights relative to each of the k topics. In this case, each row corresponds to a topic, and each column corresponds to a

unique term in the corpus vocabulary.

Topics Distributed and top words for each distributed topic is shown.



```
Topic #0:
like feel interior better feature
Topic #1:
problem dealer time issue transmission
Topic #2:
mpg highway city driving trip
Topic #3:
truck tacoma cab ram bed
Topic #4:
seat back front rear room
Topic #5:
love absolutely still look everything
Topic #6:
drive fun test comfortable lot
Topic #7:
mile tire oil change year
Topic #8:
ford focus escape fusion explorer
Topic #9:
gas mileage better get tank
```

Fig -3: Topics from topic modelling, with top words for each topic

After this stage, optimal number of topics to be keep for actual recommendation is found out from all the topics distributed. This is done to keep the recommendations from overlapping too much due to redundant topics. This is done by plotting coherence values against the number of topics.

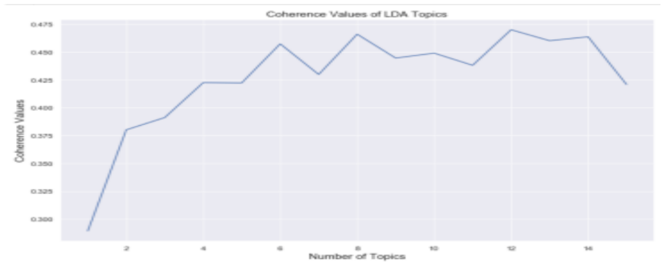


Fig -4: coherence values vs. number of topics

From the plot the number of optimal topics is determined to be 8.

d) Latent Dirichlet Allocation (LDA) model fitting: This helps us find the underlying relationships in the distributed topics so that every text pertaining to each car can be scored and each car can be assigned a topic. The topic having highest value for a particular car is assigned. And based of thi LDA model the users query is compared and the top topics for the query are scored and from that topic the top-rated cars are recommended.

The below plot shows the distribution of topics and the most salient terms and its frequency distribution.

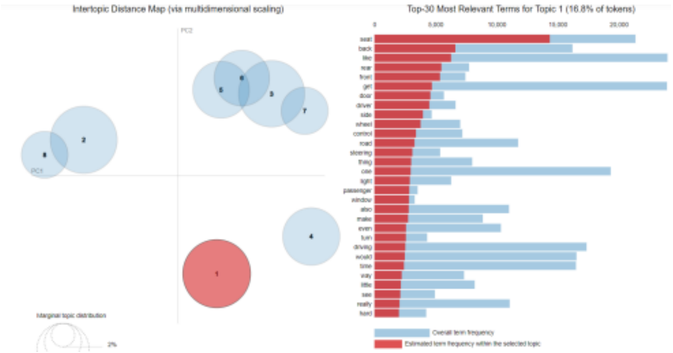


Fig -5: LDA topic distributions and word distributions for each topic

Word Count of each topic's top words from the LDA model and their Importance in that topic (Fig. 6 & Fig. 7):

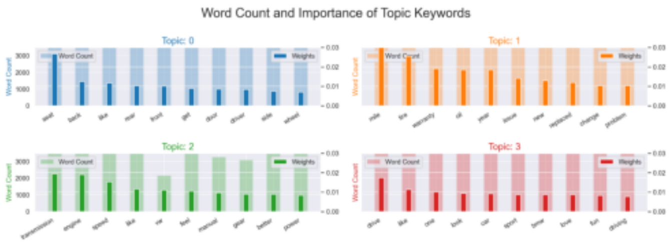


Fig -6

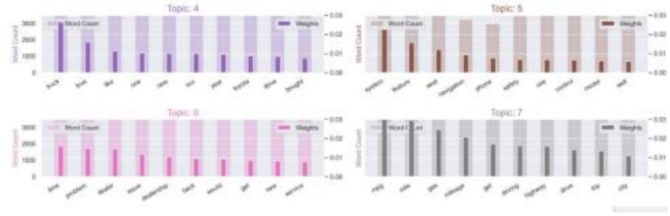


Fig -7: Visualization showing distribution of topics assigned

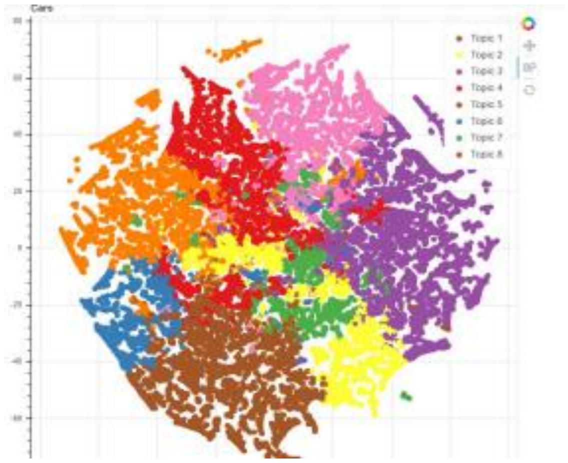


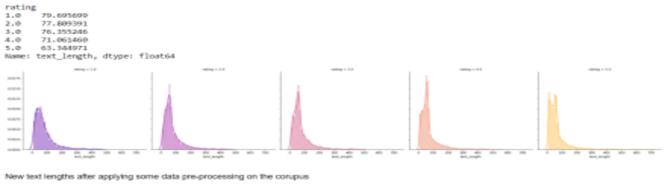
Fig -8: Visualization of topic distributed (8 optimal topics spatial distribution)

e) Recommendation: After the LDA model assigns each record with a specific topic, the users query is scored with the LDA model to find out the relations defined in the user query and find those topics which reflect the query. Then from the top topics discover the top-rated cars are recommended back to the users. For specific types of cars such as SUV and sedan or hatchback, we extracted the types of cars from their model names and appended the categories for class wise recommendation. Allowing a holistic recommendation ability based on quantitative features like category, type, etc as well as qualitative based on the users' abstract requirements.

6. RESULTS & DISCUSSIONS

Pre-processing inference:

From the visualization, it can be inferred that the ratings increase with decrease in review length, for the population.



```
rating
1.0     79.695699
2.0     77.809391
3.0     76.355286
4.0     71.061460
5.0     63.344971
Name: text_length, dtype: float64
```

New text lengths after applying some data pre-processing on the corpus

Fig -9: Preprocessing Inference

(ratings inversely proportional to review length)

The distribution shows the topics are evenly distributed and explains a different section, showing an efficient model.

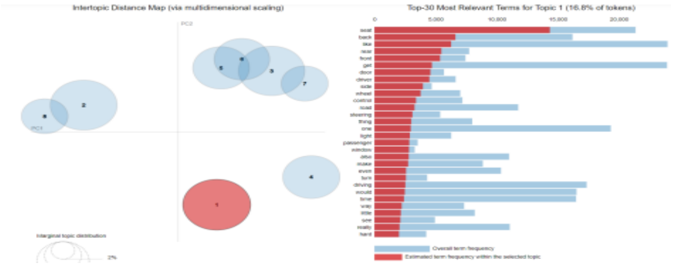


Fig -10: LDA topic distributions and word distributions for each topic

Post topic modelling: choosing optimal number of topics, from the plot the number of optimal topics is determined to be 8.

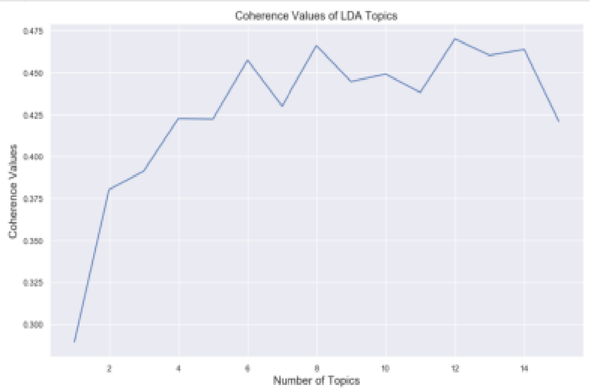


Fig -11: Coherence Values vs. No. of Topics

Topic Distributions, results: Intertopic distances show that the topics are properly spaced out and each topic explains a different category of automobile and features.

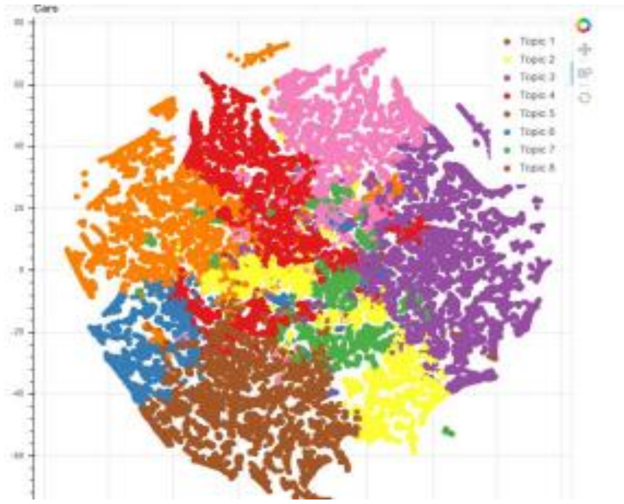
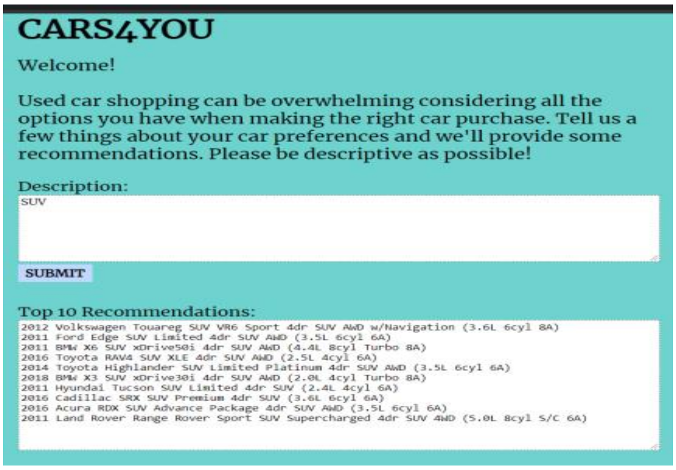


Fig -12: Visualization of topic distributed (8 optimal topics spatial distribution)

We successfully developed a flask application that that textual query as input, analyses it, runs the topic modelling and linear discriminant analysis model, and recommends top 10 cars accordingly. The recommender system uses not only quantitative inputs but also qualitative inputs which includes reviews from other customers.

Another feature that is used is that in the flask app itself, if the type of automobile is specified such as sedan or hatchback, then the query filters out those cars from the start giving better results



CARS4YOU

Welcome!

Used car shopping can be overwhelming considering all the options you have when making the right car purchase. Tell us a few things about your car preferences and we'll provide some recommendations. Please be descriptive as possible!

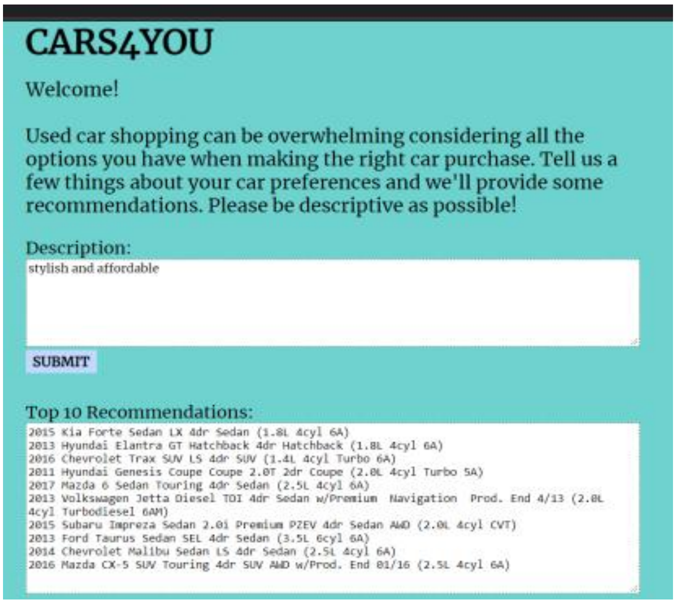
Description:

SUV

Top 10 Recommendations:

```
2012 Volkswagen Touareg SUV VR6 Sport 4dr SUV AWD w/Navigation (3.6L 6cyl 8A)
2011 Ford Edge SUV Limited 4dr SUV AWD (3.5L 6cyl 6A)
2011 BMW X6 SUV xDrive50i 4dr SUV AWD (4.4L 8cyl Turbo 8A)
2016 Toyota RAV4 SUV XLE 4dr SUV AWD (2.5L 4cyl 6A)
2014 Toyota Highlander SUV Limited Platinum 4dr SUV AWD (3.5L 6cyl 6A)
2018 BMW X3 SUV xDrive30i 4dr SUV AWD (2.0L 4cyl Turbo 8A)
2011 Hyundai Tucson SUV Limited 4dr SUV (2.4L 4cyl 6A)
2016 Cadillac SRX SUV Premium 4dr SUV (3.6L 6cyl 6A)
2016 Acura RDX SUV Advance Package 4dr SUV AWD (3.5L 6cyl 6A)
2011 Land Rover Range Rover Sport SUV Supercharged 4dr SUV 4WD (5.0L 8cyl S/C 6A)
```

Fig -13: The web application interface recommending top 10 cars for SUV type, entered by user



CARS4YOU

Welcome!

Used car shopping can be overwhelming considering all the options you have when making the right car purchase. Tell us a few things about your car preferences and we'll provide some recommendations. Please be descriptive as possible!

Description:

stylish and affordable

Top 10 Recommendations:

```
2015 Kia Forte Sedan LX 4dr Sedan (1.8L 4cyl 6A)
2013 Hyundai Elantra GT Hatchback 4dr Hatchback (1.8L 4cyl 6A)
2016 Chevrolet Trax SUV LS 4dr SUV (1.4L 4cyl Turbo 6A)
2011 Hyundai Genesis Coupe Coupe 2.0T 2dr Coupe (2.0L 4cyl Turbo 5A)
2017 Mazda 6 Sedan Touring 4dr Sedan (2.5L 4cyl 6A)
2013 Volkswagen Jetta Diesel TDI 4dr Sedan w/Premium  Navigation  Prod. End 4/13 (2.0L
4cyl Turbodiesel 6AM)
2015 Subaru Impreza Sedan 2.0i Premium PZEV 4dr Sedan AWD (2.0L 4cyl CVT)
2013 Ford Taurus Sedan SEL 4dr Sedan (3.5L 6cyl 6A)
2014 Chevrolet Malibu Sedan LS 4dr Sedan (2.5L 4cyl 6A)
2016 Mazda CX-5 SUV Touring 4dr SUV AWD w/Prod. End 01/16 (2.5L 4cyl 6A)
```

Fig -14: Recommending top 10 cars for “stylish and affordable” text query, entered by user

7. CONCLUSION

Generally, the car recommendation systems are content based, but we are trying to implement collaborative recommendation system. This will help us recommend cars using textual queries as well, which are based on the reviews given by actual customers. The existing systems are mostly either search on basis of specification or content-based recommendation.

The proposed system could be act both as specific search tool or recommender as well as an expert car advisor or a community-based suggestion, but in the digital form.

As the global market rises and demand for new products in the Indian economy is leading to the arrival of new models. All foreign car manufacturers see the Indian market as their growth point in their share of the global automotive economy. As the world progresses to the climax of a new era, recommendations become an inevitable reality. Almost all technological and non-technical items in modern hands raise their hands in compliments. The main fact that the recommendations are extremely focused on the new technology is because of its accuracy, precision, and reliability.

The recommendation provides a personal preference for user needs. In the proposed method, which is a combination of user-to-user and object to a collaboratively based object to recommend filtering an algorithm that works well for suggesting. The biggest problem with car databases is that they are dynamic data because it is difficult to predict the car model that will be released in their product. In addition, the performance of the proposed system can be improved by using a real-time network that allows you to build websites and access session information. This research activity can be expanded as information-based complimentary programs using a variety of information presentations. Expert recommendations using a professional program can also be considered using knowledge bases.

REFERENCES

[1] Dataset sources, Edmunds.com / Kaggle.com,

(https://www.kaggle.com/ankkur13/edmundsconsumer-car-ratings-and-reviews)

[2] Q. Zhang, J. Wang and K. Fan, "Research on passenger car recommendation based on comments mining of Internet," 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2017, pp. 122-127, doi: 10.1109/ICIEA.2017.8282826.

[3] G. Prabowol, M. Nasrun and R. A. Nugrahaeni, "Recommendations for Car Selection System Using

Item-Based Collaborative Filtering (CF)," 2019 IEEE International Conference on Signals and Systems (ICSigSys), 2019, pp. 116-119, doi: 10.1109/ICSIGSYS.2019.8811083.

[4] Shrey Talati, Anukrity, Priyanka Salian and Anam Hussain. Article: Recommendation System for Automobile Purchasing: A Survey. IJCA Proceedings on National Conference on Advancements in Computer & Information Technology NCACIT 2016(6):23-27, May 2016.

[5] Vengatesan K, A. Srivastava, A. Kumar, S. Samee, P. T. Vijay, P. S. Tanesh, “A Novel Approach of Car Recommendation Using Machine Learning Algorithm”, (https://www.ijstr.org/final-print/jun2020/A-Novel-Approach Of-Car-Recommendation-Using-Machine-Learning-Algorithm.pdf)