Sentiment Analysis via applied machine learning algorithms from Honda review Data

Md Rakibul Hasan[a], Miraziz Salehin[b], Mohammad Mansur Billah[c], Khandaker Tanvir Ahmed[d], Afsana Khondokar Mimi[e]

rakibul15-11826@diu.edu.bd, Miraziz15-11901@diu.edu.bd, mansur15-11899@diu.edu.bd, Khandaker15-11923@diu.edu.bd, afsana15-11922@diu.edu.bd

[a] Daffodil International University, Dhaka -1212, Bangladesh
[b] Daffodil International University, Dhaka -1212, Bangladesh
[c] Daffodil International University, Dhaka -1212, Bangladesh
[d] Daffodil International University, Dhaka -1212, Bangladesh
[e] Daffodil International University, Dhaka -1212, Bangladesh

Abstract

The reason behind this analytic debris is to evaluate the review data about the motorcycle Honda unicorn 150cc.Honda motor company ltd. is a Japanese public transport manufacturing organization which has their headquarters located in Minato, Tokyo, Japan. This work is allocated to contradict the review which is based on real time comments taken from the review session. Opinion mining or Emotion Artificial intelligence is a present formed field where the machine learning algorithms are used for the analysis of the information gathered from people's real time opinions. The purpose behind this work is to predict the given motorcycles user satisfaction and negative opinions. The collected from "mouthshut.com". Then the dataset was optimized and manually labeled and optimized for further analysis procedures. So, by our analysis we have accomplished our results about public inspections and comments which represents their thoughts on the model's basics, this result of ours can be used in the sole purpose the company's upcoming motorcycle upgrades, modifications and overcoming motorcycles predicaments venture.

Keywords: Sentiment Analysis; Machine Learning; Artificial Intelligence; Review Analysis

1. Introduction

The motorcycle we have chosen is Honda unicorn 150cc.This motorcycle was launched by the Honda company in 2004 [1]. The motorcycle's engine capacity is 162.7 cc in total. Its average mileage is around 50 kmpl. The transmissions are 5 speed manuals. The kerb weight is around 140kg and the fuel tank capacity is 13 liters. The seat length is 798mm [2]. The motorcycle is available for purchasing only in one variant and three different colors. The looks of the motorcycle are kind of classic. The motorcycle is currently available in Asian countries. It's a very popular motorcycle because of its low price and good performance. The data that we worked on was labeled with three different classifications, which are Positive, Negative and Neutral. Then the pre-processed dataset was analyzed using seven very profound algorithms which are Logistic regression, KNeighborsClassifier, GaussianNB, Decision tree Classifier, Random forest Classifier, SVC, XGB. The

algorithms have provided a final outcome result which was 98%. Which was pretty acceptable when it comes to sentiment analysis section.

2. Experiments and Methods

For this work we have collected the dataset from the reviews from "mouthshut.com", in the dataset we genuinely used the machine learning algorithms for an accurate and better outcome, The dataset contains the total number of 5009 data's which were manually labeled. In the dataset the three classes are divided in numbers.

The data was in the string form which and why this data needed to be optimized into numerical validation for our model. The data were utilized for the elimination of the useless or compromised data. The dataset was divided into two portions which are test 20% and train 80%.

TABLE 1. DATASET

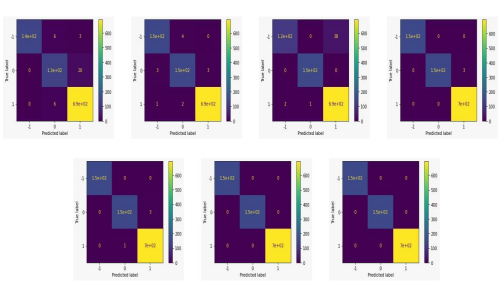
Class	*Amount*
Positive	*3548*
Negative	*757*
Neutral	*704*

TABLE 2. VARIABLE CLASSIFICATION

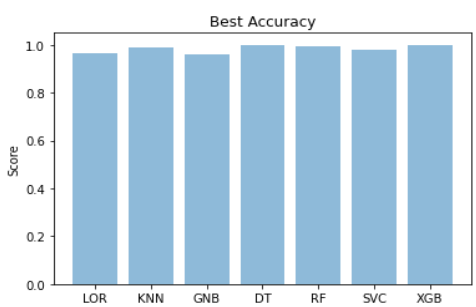
Variable	Class	Training Dataset	Testing Dataset
01	Binary (1,0,1)	4008	1001

3. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.



(a)



(b)

Figure 1. (a) Description of what is contained in the first panel; (b) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centred.

TABLE 3. CLASSIFICATION REPORT

Algorithm	*Class*	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Logistic Regression	*-1*	*0.97*	*0.96*	*0.94*	*0.97*
	0	*0.96*	*0.94*	*0.97*	
	1	*0.94*	*0.94*	*0.96*	
KNeigbours	*-1*	*0.96*	*0.96*	*0.97*	*0.97*
	0	*0.94*	*0.94*	*0.96*	
	1	*0.97*	*0.97*	*0.97*	
GaussianNB	*-1*	*0.96*	*0.94*	*0.95*	*0.96*
	0	*0.95*	*0.99*	*0.96*	
	1	*0.95*	*0.96*	*0.96*	
Decision	*-1*	*0.97*	*0.98*	*0.96*	*0.98*
	0	*0.97*	*0.96*	*0.94*	
	1	*0.96*	*0.94*	*0.97*	
Random	*-1*	*0.94*	*0.94*	*0.96*	*0.95*
	0	*0.96*	*0.96*	*0.97*	
	1	*0.94*	*0.94*	*0.96*	
SVC	*-1*	*0.97*	*0.97*	*0.97*	*0.97*
	0	*0.96*	*0.94*	*0.95*	
	1	*0.95*	*0.99*	*0.96*	
XGBoost	*-1*	*0.95*	*0.96*	*0.96*	*0.96*
	0	*0.97*	*0.98*	*0.96*	
	1	*0.97*	*0.96*	*0.94*	

Then we applied algorithms deliberately, the descriptions of these algorithms are given below,

1.Logistic Regression: The logistic regression is mainly used when the dependent variables are binary. The main purpose of this algorithm is to explain data and establish relation between the binary variables and the other variables. The algorithm represents in 0 and 1.

2.KNeighborsClassifier: This algorithm is for the classification and regression in the model. The algorithm does its part of getting better accuracy and calibration of the data's.

3.GaussianNB: The GaussianNB is mainly a version of the NB algorithm. This algorithm is mainly used on large datasets that have continuous evaluation.

4.Decision tree Classifier: The decision tree has the part of creating a tree that can classify that data's that are converted into 0 and 1. It can also predict the outcome result. It is very known among the machine learning data analysis works.

5.Random Forest Classifier: Random Forest classifier has the similar credentials as the decision tree. It also creates similar tree segments for creating the binary outcome and more accurate performance.

6.SVC: SVC is mainly a version of support vector machine, which has mainly used in real life-based problems.
7.XGM: This algorithm works with structured data's such as our works. This is also a very familiar algorithm in machine learning algorithm structure.

4. Discussion

The main intention of this paper is to check which algorithm performed at its best. As we all know the concept that not all algorithms work in the same way when it comes to raw dataset. So, we wanted to know which of our algorithms will work in our case. So, in that case we were willing to determine this motorcycle's review within the best result output detected by our preferred algorithm, which helped us to confirm the sentiment conception of the motorbike Honda unicorn 150cc performance.

5. Conclusions

The result we achieved shows which algorithm gave the best performance from our dataset analysis. The highest accuracy we gathered was from decision tree 0.98%. In such a case we were able to determine the best accuracy from our motorbikes comment dataset and able to recollect about the bike's performance evaluation between the good, neutral and bad comments. Which can help us to evaluate other review datasets in the future.

Acknowledgments: We would like to express our gratitude to our PetaAI Solutions for providing all research resources. Our supervisor deserves special thanks for his patience and support in helping us overcome the lot of obstacles we encountered during our research.

References

Honda Unicorn - Wikipedia. (2021). Retrieved 16 July 2021, from https://en.wikipedia.org/wiki/Honda_Unicorn

Chumwatana, T. (2018). COMMENT ANALYSIS FOR PRODUCT AND SERVICE SATISFACTION FROM THAI CUSTOMERS'REVIEW IN SOCIAL NETWORK. Journal of Information and Communication Technology, 17(2), 271-289.

Liu, L., Wang, Y. H., & Wei, H. (2015). Fine-grained sentiment analysis oriented to product comment. Journal of Computer Applications, 35(12), 3481-3486.

McArdle, B. H., & Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance‐based redundancy analysis. Ecology, 82(1), 290-297.



Md. Rakibul Hasan, was born Lalmonirhat, Sadar, Bangladesh on December 13 March, 1999. Now he is studying Computer Science and Engineering at Daffodil International University, Dhaka, Bangladesh. He is currently working on research related. Gaining knowledge about data mining, machine learning also interested in working broadly in the deep learning process. Besides, now, he is serving as a CAO in petaAI Solutions, Dhaka, Bangladesh.



Miraziz Salehin, was born Lakshmipur, Bangladesh on November 7, 1999. Now he is studying Computer Science and Engineering at Daffodil International University, Dhaka, Bangladesh. He is currently working on research related. Gaining knowledge about data mining, machine learning also interested in working broadly in the deep learning process. Besides, now, he is serving as a CTO in petaAI Solutions, Dhaka, Bangladesh.



Mansur Billah, was born Hatiya, Bangladesh on November 18, 1999. Now he is studying Computer Science and Engineering at Daffodil International University, Dhaka, Bangladesh. He is currently working on research related. Gaining knowledge about data mining, machine learning also interested in working broadly in the deep learning process. Besides, now, he is serving as a CPO in petaAI Solutions, Dhaka, Bangladesh



Khandaker Tanvir Ahmed, was born Tangail, Dhaka, Bangladesh on December 1, 1998. Now he is studying Computer Science and Engineering at Daffodil International University, Dhaka, Bangladesh. He is currently working on research related. Gaining knowledge about data mining, machine learning also interested in working broadly in the deep learning process. Besides, now, he is serving as a HR in petaAI Solutions, Dhaka, Bangladesh.



Afsana Khandokar Meme, was born Tangail, Dhaka, Bangladesh on October 25, 2000. Now he is studying Computer Science and Engineering at Daffodil International University, Dhaka, Bangladesh. He is currently working on research related. Gaining knowledge about data mining, machine learning also interested in working broadly in the deep learning process. Besides, now, he is serving as a COO in petaAI Solutions, Dhaka, Bangladesh.