iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51918

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Analysis on Automobile Reviews using Machine Learning Techniques

Dr. B. Meena Preethi[1], Ms. G. P. Kavyasri[2], Ms. V. S. Akshaya[3]
[1]*Assistant Professor-Department Of Computer Science,* [2, 3]*B.Sc.Data Science - Department Of Computer Science, Sri Krishna Arts and Science College*

Abstract: ***Automobile reviews provide a valuable opportunity for sentiment analysis in order to gain a better understanding of consumer opinions and improve product quality. As a result of the complexity of natural language and differences in user opinions, accurate and robust sentiment analysis by computer models remains challenging. Traditional machine learning algorithms such as Support Vector Machines (SVM) and Naive Bayes (NB) have been used for sentiment analysis on textual data. They can, however, be limited by the heterogeneity of user opinions and language variations. In contrast, deep learning models and natural language processing have shown great promise in sentiment analysis. In particular, the VADER (Valence Aware Dictionary and sentiment Reasoner) algorithm has been proposed as a powerful tool for sentiment analysis on social media texts, and has shown promising results in other domains as well. In this work, The aim is to achieve accurate sentiment analysis on automobile reviews by comparing the accuracy performance of SVM, NB, and the VADER algorithm.***
Keywords: (SVM, NB, VADER, TOKENIZATION, STEMMING, MINNING)

I. INTRODUCTION

Sentiment analysis, also known as opinion mining, is a computational technique used to identify and extract subjective information from text data. In recent years, sentiment analysis has gained significant attention in various fields such as marketing, customer service, and product development, among others. In the automotive industry, analyzing the sentiment of customer reviews is crucial for understanding customer satisfaction and identifying areas for improvement.

In this study, the aim is to perform sentiment analysis on automobile reviews using three different methods: Support Vector Machines (SVM), Naive Bayes (NB), and VADER. SVM and NB are traditional machine learning algorithms widely used in text classification tasks. VADER, on the other hand, is a rule-based sentiment analysis tool that has gained popularity due to its high accuracy and speed. A dataset of automobile reviews collected from online sources will be used. The goal is to train the SVM and NB models on this dataset and evaluate their performance in classifying the reviews as positive or negative. VADER will also be used to perform sentiment analysis and compare its performance with the traditional machine learning models.

Through the experiments, the aim is to identify the most effective method for sentiment analysis on automobile reviews and provide insights into the sentiment of customers towards different aspects of automobiles.

II. PROBLEM DEFINITION

The popularity of online platforms for car reviews and opinions has made sentimental analysis on automobile reviews an important task. Early identification of negative sentiment towards certain aspects of a vehicle can help manufacturers improve their products and services, leading to increased customer satisfaction and loyalty. The existing system of this project uses support vector machine and naïve Bayes, which have shown lower accuracy. Therefore, a proposed model aims for higher accuracy using the Vader algorithm. VADER is a rule-based sentiment analysis tool that uses a lexicon of words and their associated sentiment scores to determine the sentiment of a piece of text.

III. LITERATURE SURVEY

Sara Ashour Aljuhani and Norah Saleh Alghamdi[1] has done research paper on “comparison of sentimental analysis method on amazon phone reviews” where various algorithms Regression (LR), Naive Bayes (NB), Stochastic Gradient Decent (SGD) and deep learning algorithms such as CNN. These algorithms are applied using different feature extraction approaches. The objective of this study was to compare the performance of different sentiment analysis methods on Amazon reviews of mobile phones. Dataset: The authors used a dataset of 1600 Amazon reviews of mobile phones, which were manually labelled as positive, negative, or neutral. The authors reported that SVM achieved the highest accuracy of 84.3%, followed by Naive Bayes (80.6%), VADER (76.4%), and TextBlob (75.8%). They also reported the precision, recall, and F1-score for each class

Smita Bhanap and Dr. Seema Babrekar[2] studied the "Feature Selection and Polarity Classification using Machine Learning Algorithms NB & SVM " and also found the accuracy for most positive rated mobile phones. The objective of this study was to perform sentiment analysis of product reviews using machine learning algorithms (NB and SVM) and feature selection techniques. Dataset: The authors used a dataset of 1000 product reviews from Amazon, which were labelled as positive or negative. The authors reported that the SVM algorithm achieved the highest accuracy of 81.4%, followed by Naive Bayes (80.6%). They also reported the precision, recall, and F1-score for each class, and found that SVM achieved high scores for all three metrics.

Chaithra V. D[3] Proposed a "hybrid method approach on analysing naive bayes and sentiment VADER " for sentiment of mobile unboxing video comment. The objective of this study was to analyse the sentiment of mobile unboxing video comments using a hybrid approach of Naive Bayes and sentiment VADER algorithms. The author used a dataset of 1000 comments from mobile unboxing videos on YouTubes. The author reported that the hybrid approach achieved an accuracy of 84.3%. They also reported the precision, recall, and F1-score for each class, and found that the hybrid approach achieved high scores for all three metrics.

Mohamed Chiny , Marouane Chihab , Younes Chihab [4] Proposed LSTM, VADER and TF-IDF based Hybrid Sentiment Analysis Model on IMDB Dataset.The objective of the paper is to propose a hybrid model that can improve the accuracy and generalization of sentiment analysis models. The authors aimed to overcome the limitations of using a single algorithm for sentiment analysis by combining three algorithms - LSTM, VADER, and TF-IDF.The authors used the IMDB dataset, which is a popular dataset for sentiment analysis. The dataset consists of 50,000 movie reviews, with 25,000 reviews for training and 25,000 for testing. Each review is labeled as either positive or negative. The proposed hybrid model achieved an accuracy of 92.86%, which was higher than the accuracy achieved by using individual algorithms.

R. V. Vidhate and N. R. Jagdale[5]"Comparative Study of Sentiment Analysis of Automobile Reviews using Various Techniques"(2018) Algorithm used: The authors compared the performance of various sentiment analysis techniques, including VADER, Naive Bayes, and Support Vector Machine. The VADER algorithm achieved an accuracy of 90.4%, which was lower than the accuracy achieved by Naive Bayes and Support Vector Machine-(93.8%).

Deepthi V. and Sudarshan P[6] "Sentiment Analysis of Automobile Reviews using VADER Algorithm"*(2019) The objective of this study was to perform sentiment analysis on automobile reviews and evaluate the performance of the VADER algorithm for this task.The authors used a dataset of 1000 automobile reviews collected from a popular automobile review website. The authors reported an accuracy of 91.5% for the VADER algorithm in classifying the automobile reviews as positive, negative, or neutral. They also reported the precision, recall, and F1-score for each class, and found that the VADER algorithm achieved high scores for all three metrics.

L. V. R. Koteswari and K. Ramesh[7] "Sentiment Analysis of Car Reviews using Machine Learning Algorithms and VADER"(2020) The objective of this study was to perform sentiment analysis on car reviews and compare the performance of machine learning algorithms (SVM and NB) and the VADER algorithm for this task. The authors reported that the combination of machine learning algorithms and VADER achieved the highest accuracy of 95.5% in classifying the car reviews into positive, negative, or neutral categories. The SVM algorithm achieved an accuracy of 94.5%, while the NB algorithm achieved an accuracy of 93.5.

IV. METHODOLOGY

The methodology and techniques used in classifying reviews are commonly adopted by researchers in the field of sentiment analysis. The steps followed during the experiments will be explained, starting with the review dataset and ending with the classification of each review as positive or negative. Fig. 1 illustrates the phases of this work

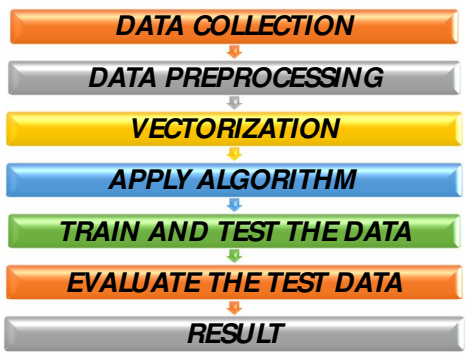


FIG 1: Sentiment Analysis Process

A. *Data Collection*

Data collection is the process of collecting accurate insights for research. In most cases, data collection is the primary step for research. In our study, dataset is taken from Kaggle which contains ford company reviews with the size of 20717 rows & 7 columns with more than 20000 reviews and presented as follows (Date, author name, vehicle title, review_title,review,rating).

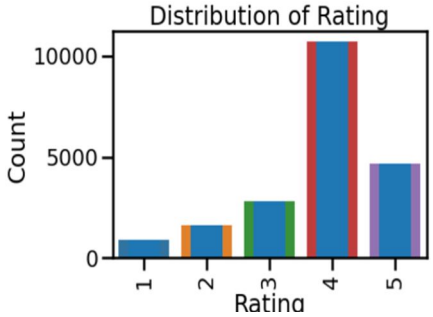


FIG 2: Total count of Rating in the automobile dataset

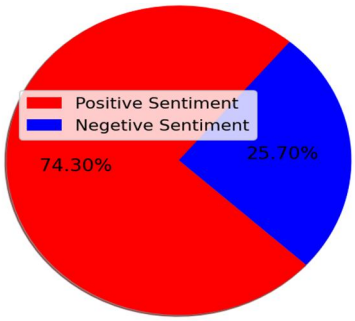


FIG 3: Total no of positive & negative reviews in the dataset

B. *Data Pre-Processing*

Pre-processing is an essential step in sentiment analysis that involves cleaning and transforming raw data into a format suitable for analysis. The pre-processing step aims to reduce noise and ensure that the data is in a consistent and normalized format for analysis. Here are some common pre-processing steps in sentiment analysis:

1) *Text Cleaning:* The first step in pre-processing is to clean the text data by removing unwanted characters, punctuation, and special characters. This can be done using regular expressions or pre-built libraries like NLTK or spaCy.
2) *Tokenization:* Once the text has been cleaned, the next step is to tokenize the text into individual words. Tokenization involves breaking down the text into smaller parts (tokens) that can be easily analysed. This step is crucial as it helps to identify the sentiment-bearing words in the text.
3) *Stop Words Removal:* Stop words are common words that do not carry much meaning in the text, such as "a," "the," "and," etc. These words can be removed to reduce the noise in the text data and focus on the important words that carry sentiment.
4) *Stemming and Lemmatization:* These techniques involve reducing words to their root form to capture their essence. Stemming involves removing the suffixes of words, while lemmatization involves mapping words to their base form. These techniques are used to reduce the dimensionality of the data and improve the accuracy of the sentimental analysis.

Table 1: Example of Pre-processing Stage

Before pre-processing	There will always be a 05-09 mustang for sale and their reasonable. Purchased mine as second car and I believe it was a great investment
After pre-processing	Mustang,Sale,Fairly, Reasonable , Purchased, Great Investment

V. VECTORIZATION

In sentiment analysis, vectorization plays a critical role in turning raw text data into features that can be used to classify the sentiment of a particular text. There are several vectorization techniques used in sentiment analysis, including: Bag of Words (BoW): This technique involves creating a vocabulary of all the words in the dataset and then representing each document as a vector of word frequencies. Term Frequency-Inverse Document Frequency (TF-IDF): This technique is similar to BoW, but it also takes into account the rarity of words in the dataset by weighting them based on how often they appear in the document and the overall dataset. Word Embeddings: This technique represents words as dense vectors in a high-dimensional space, where semantically similar words are closer together. Word embeddings are often pre-trained on large corpora of text and can be used to transfer learning to new sentiment analysis tasks. Vectorization is essential in sentiment analysis because it allows machine learning models to process and classify text data accurately. By representing textual data as numerical vectors, machine learning models can use mathematical techniques to analyse and classify sentiment in large datasets. This paper is mainly focused on Count vectorizer.

Count Vectorizer is a vectorization technique used in natural language processing and text analytics to convert a collection of text documents into a matrix of token counts. It is a method of representing text data numerically, and it is commonly used in sentiment analysis. In Count Vectorizer, each document is represented as a vector of word counts. The rows of the matrix represent the documents in the corpus, and the columns represent the individual words or tokens. The values in the matrix represent the frequency of each word in each document. For example, assume it have a corpus of three documents:

"I love car"

"Car is my favorite vehicle."

"I drive car every day."

Using Count Vectorizer, represent this corpus as a matrix of token counts:

TABLE 2: Matrix Of Token Counts

	drive	every	favorite	vehicle	love	car
D1	0	0	0	0	1	1
D2	0	0	1	1	0	1
D3	1	1	0	0	0	1

The Count Vectorizer technique can be applied in Python using the scikit-learn library. It allows for easy pre-processing of text data, including tokenization, stop word removal, and stemming. Once the text data is pre-processed, the Count Vectorizer can be fit to the data to create the matrix of token counts. Count Vectorizer is a simple yet powerful vectorization technique that can be used in a wide range of natural language processing tasks, including sentiment analysis, text classification, and topic modelling.

VI. ALGORITHMS

Support Vector Machine (SVM) is a machine learning algorithm used in sentiment analysis to classify text into positive, negative, or neutral sentiment. SVM works by finding the optimal hyperplane that separates the different classes of sentiment. It is a popular algorithm for sentiment analysis because it is effective at handling high-dimensional data and has a good generalization performance. Naive Bayes is another popular algorithm used in sentiment analysis. It is a probabilistic machine learning algorithm that works by calculating the probability of a text belonging to each sentiment class. The algorithm assumes that each feature (word) in the text is independent of all other features, which is why it is called "naive." Despite its simplicity, Naive Bayes can be very effective in sentiment analysis and is often used as a baseline algorithm. VADER algorithm is a rule-based sentiment analysis tool that uses a lexicon of words and their sentiment scores to calculate the overall sentiment of a piece of text. VADER is designed to handle social media text, which often contains slang, sarcasm, and other forms of language that are not well-handled by other sentiment analysis algorithms.

VII. CLASSIFICATION

The Train-Test-Split technique was used to divide the dataset into 80% for training and 20% for testing. Then applied different classification algorithms of Supervised Machine Learning on training data to train Machine Learning Classifiers and tested with testing data. Applied algorithms are Support Vector Machine, Multinomial Naive Bayes and lexicon-based approach method: Vader.

In addition to overall sentiment classification, it is also possible to classify reviews based on the most positive rate vehicle. This involves identifying the vehicle that has received the highest number of positive reviews and then analysing the sentiment of those reviews to understand the reasons behind the positive sentiment. This can help automobile manufacturers and dealers to identify the strengths of their vehicles and use that information to improve their marketing and sales strategies.

VIII. RESULTS AND DISCUSSION

Our experimental results are shown in the below figure as a graphical representation. Table shows results of SVM classifier and VADER classifier. By the results acquired it is showing that the Vader algorithm has received more accuracy compared to other two algorithm Support Vector Machine and Naïve Bayes.

Table 3
Comparison of algorithms Accuracy

ALGORITHM	ACCURACY
SVM	50
NB	56
VADER	80

TABLE 4
Comparison of algorithms Accuracy on most positive rated vehicle

ALGORITHM	ACCURACY
SVM	64
NB	35
VADER	85

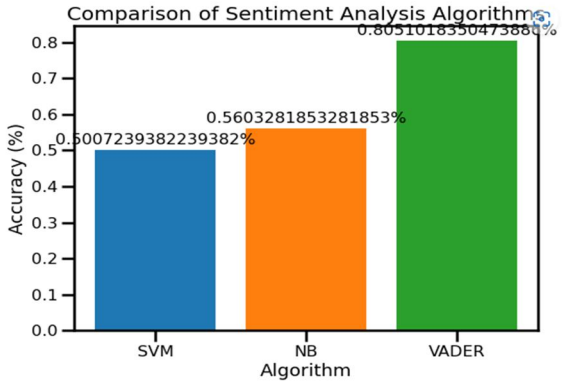


FIG 4 Comparison of algorithms

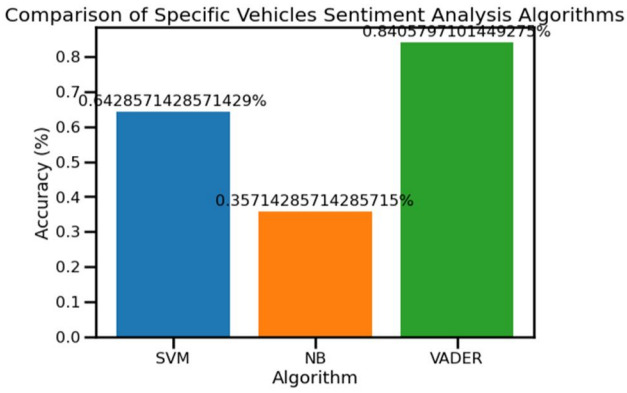


FIG 5 Comparison of Algorithms Accuracy on Most Positive Rated Vehicle

IX. CONCLUSION AND FUTURE ENHANCEMENT

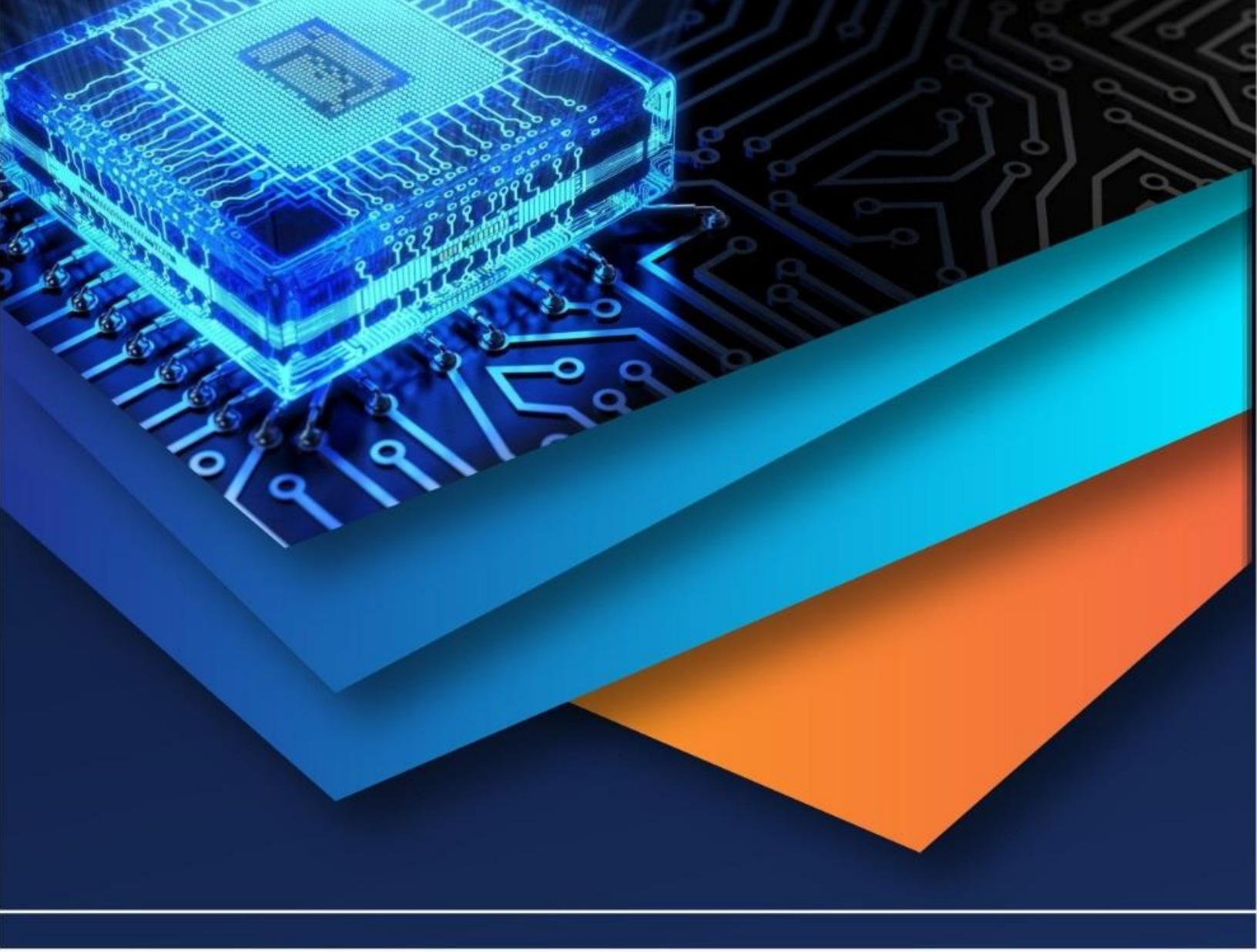
In this project, Sentimental analysis was performed on automobile reviews using three algorithms: SVM, Naive Bayes, and VADER. The data was pre-processed using stemming and stop words removal, and features were extracted using the Count Vectorizer. The data was then split into training and testing sets, and the models were trained. The accuracy of Vader was found to be the highest among all three models.

In terms of future enhancements, it would be worth exploring other feature extraction techniques such as TF-IDF or Word2Vec and experimenting with other algorithms like Random Forest, Gradient Boosting, or Neural Networks. Additionally, more advanced techniques such as deep learning and natural language processing could also be used to improve the accuracy of sentimental analysis.

In conclusion, sentimental analysis is a valuable tool for businesses to understand customer sentiment and improve their products and services. SVM, Naive Bayes, and VADER are effective algorithms for sentimental analysis, with each having its strengths and weaknesses. The choice of algorithm ultimately depends on the specific use case and the nature of the data.

REFERENCES

[1] Sara Ashour Aljuhani, Norah Saleh Alghamdi "A Comparison of Sentiment Analysis Methods on Amazon Reviews of Mobile Phones" , International Journal of Advanced Computer Science and Applications,Vol. 10, No. 6, 2019
[2] Smita Bhanap, Dr. Seema Babrekar "feature Selection and Polarity Classification using Machine Learning Algorithms NB & SVM "International Conference on Communication and Information Processing (ICCIP-2019)
[3] Chaithra V. D. "Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments "International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 5, October 2019.
[4] Mohamed Chiny , Marouane Chihab , Younes Chihab , Omar Bencharef "LSTM, VADER and TF-IDF based Hybrid Sentiment Analysis Model" International Journal of Advanced Computer Science and Applications, Vol. 12, No. 7, 2021
[5] R. V. Vidhate,N. R. Jagdale."Comparative Study of Sentiment Analysis of Automobile Reviews using Various Techniques"(2018).
[6] Deepthi V, Sudarshan"Sentiment Analysis of Automobile Reviews using VADER Algorithm"*(2019).
[7] L. V. R. Koteswari, K. Ramesh "Sentiment Analysis of Car Reviews using Machine Learning Algorithms and VADER" (2020)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY