IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 3, March 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104185 |

E-Commerce Products Reviews using Data Mining

Chitra Atlani [1] , Nikita Narwani [2], Navin Idnani [3] , Tamanna Bhatija[4] , Dr. Dashrath Mane[5]

U.G. Student, Department of Computer Engineering, Vivekanad Education Society Institute of Technology College, Chembur, Maharashtra, India[1,2,3,4]

Assistant Professor, Department of Computer Engineering, Vivekanand Education Society Institute of Technology College, Chembur, Maharashtra, India[5]

ABSTRACT: Sentiment analysis is the technique of using data mining to study, view, or analyze sentences in order to anticipate their emotional content. It divides the text into three categories and assigns positive and negative labels to the "better" and "worse" sentiments. The World Wide Web (WWW) has become a vast source of user- or custom-generated raw data, and sentiment analysis has emerged to automatically analyze such massive amounts of data. This paper proposes a novel strategy and surveys several sentiment analysis methods to determine the polarity or emotion of a user or client.

KEYWORDS: Sentiment Analysis, Naïve Bayes, Mining, Support Vector Machine, Polarity, Semantic, Accuracy.

I. INTRODUCTION

E-commerce product reviews are a trusted source of information for consumers, but not all reviews are created equal. Positive reviews can increase the perceived value of a product and lead to increased sales, while negative reviews can hurt a product's reputation and lower sales. Understanding consumer perceptions is important for businesses and marketers looking to succeed in the online marketplace. Online reviews are essential for businesses to boost sales and improve their standing, evaluation, and reputation. Product ratings are also a consideration for consumers when making an online purchase.Sentiment analysis is an important task for businesses to understand customer feedback and improve their products and services accordingly. Product review websites have been scrapped by Amazon, Flipkart, and Walmart.

II. RELATED WORK

Manek et al. proposed feature extraction methods, Hai et al. proposed SJSM, Singh et al. performed text sentiment analysis, Huang et al. proposed a multi-modal joint sentiment theme model, and Huq et al. used SVM and KNN to analyze Twitter sentiment.

Sentiment Analysis for Social Images is a state-of-the-art model for sentiment analysis on social photos using transfer learning techniques. FeD outperforms SVM and sLDA in terms of performance by 1.08-1.18. Apoorv Agarwal and Vivek Sharma use SentiWordNet to do opinion mining on news headlines.

Sentiment Analysis for Social Images is a state-of-the-art model for sentiment analysis on social photos. FeD outperforms SVM and sLDA in terms of performance by 1.08-1.18. Ontology-based Aspect Extraction for Better Sentiment Analysis in Product Review Summarization is an ontology-based sentiment summarization framework that performs better than other existing methods.

III. METHODOLOGY

A. Data Collection and Acquisition

1. Web scraping is the process of automatically extracting information from websites using a program or script. Python is a popular language for web scraping due to its ease of use, wide range of libraries, and powerful data manipulation capabilities. Beautiful Soup is a Python library for web scraping and parsing HTML and XML documents, providing easy navigation, flexibility, and integration with other libraries.
2. Handling Missing Values:Missing values in data sets can lead to biased results if not handled appropriately. Approaches such as exclusion, imputation, or modern statistical methods can help to address the issue. Our

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104185 |

project used graphs to visualize and analyze missing data and compare it to other methods.

B. Data Preprocessing

Text-Cleaning:

1. Removing punctuation and special characters: This involves removing all non-alphabetic and non-numeric characters from the text, such as commas, quotes, brackets, and dashes.
2. Tokenization: This is the process of splitting the text into individual words or tokens. This is an important step in natural language processing (NLP), as it allows us to analyze text at the word level.
3. Stop word removal: Stop words are words that occur frequently in a language but do not carry much meaning, such as "the", "and", and "in". Removing stop words can help to reduce the dimensionality of the text and improve the accuracy of text analysis.
4. Stemming and lemmatizationThese techniques involve reducing words to their base form to reduce the dimensionality of the text and improve the accuracy of sentiment analysis. Stemming involves removing the suffixes and prefixes from a word to reduce it to its root form, while lemmatization involves converting a word to its dictionary form based on its part of speech. Ultimately, the choice between stemming and lemmatization depends on the specific needs of the project and the nature of the text data. Stemming may be more appropriate for applications where speed and simplicity are important, while lemmatization may be more appropriate for applications where accuracy and context are important. But for our datasets of Flipkart we have found stemming to be more appropriate as it gave better results.
    a. In our project, we have found that steaming is better and more accurate for datasets
5. Spell checking and correction: This involves identifying and correcting spelling errors in the text. This can help to improve the accuracy of text analysis, particularly for tasks such as sentiment analysis or topic modeling.
6. Emoticon remover: Letters, punctuation, and numerals are used in combination to represent facial expressions of emotion. People frequently use emoticons to convey their moods. Emoticon remover is used for cleaning purposes to modify sentences.

Overall, text cleaning is an important preprocessing step in sentiment analysis of e-commerce product reviews. By removing irrelevant or redundant information from the text, and transforming the text into a format that is suitable for analysis, text cleaning can help to improve the accuracy of sentiment analysis and provide valuable insights into customer opinions and preferences.

C. Feature extraction

For customer reviews sentiment analysis is the process of identifying the most relevant and informative features or attributes of a product that are mentioned in customer reviews and using them to train a machine learning model to predict the sentiment of new reviews. The first step is collecting the dataset and pre-processing it, followed by identifying the most relevant and informative features. The features are then extracted and represented in a way that can be used by a machine learning algorithm. The accuracy of the model depends on the quality of the feature extraction and representation, as well as the choice of the machine learning algorithm and its parameters.

D. Classifier data Algorithms(Model Building)

- MultiNomial Naive Bayes Classifier

Multinomial Naive Bayes (MNB) is a classification algorithm based on Bayes' theorem with an assumption of the multinomial distribution of features. It is commonly used for text classification tasks, such as sentiment analysis or topic classification, and works by calculating the probability of each class given the features of the input and selecting the class with the highest probability as the output. The formula for MNB can be expressed as P(class | features) = P(class) * product(P(feature_i | class)). MNB is particularly suited to text classification tasks because it can handle large numbers of features and still make accurate predictions with relatively small training datasets. However, it may struggle with rare or unseen features in the test data, as it relies on counting occurrences of each feature in the training set.

- Support Vector Machine

Support Vector Machines (SVM) is a popular machine learning algorithm used for classification and regression tasks. The goal of SVM is to find the best hyperplane that separates the data into different classes. The basic formula for SVM can be expressed as follows: wT x + b = 0 where w is the normal vector to the hyperplane and b is the bias term. The goal of SVM is to find the best hyperplane that maximizes the margin between the two classes. The optimization problem for SVM can be formulated as follows: minimize: 1/2 ‖w‖2 subjects to yi (wT xi + b) >= 1 This means we

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104185 |

want to find the minimum value of ||w||2 while ensuring that all data points are correctly classified by the hyperplane.

The SVM algorithm is used to solve various optimization techniques, such as quadratic programming or gradient descent. Once the hyperplane is found, it can be used to predict the class of new data points.

- Logistic Regression

It is a popular machine-learning algorithm that can be used for sentiment analysis. The algorithm is a type of binary classification model that can predict whether a piece of text has a positive or negative sentiment. Here's a brief overview of how logistic regression can be used for sentiment analysis:
It is a simple and interpretable algorithm that can work well for sentiment analysis tasks. However, a binary classification model may not be suitable for tasks that require multi-classification algorithms.

- Random Forest Classifier

Random Forest is a popular machine learning algorithm used for both classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to make predictions. It has several advantages over a single decision tree, such as reduced overfitting, improved accuracy, robustness, and feature importance. It can be used for both classification and regression tasks, assigning the class that receives the most votes from the individual decision trees, and calculating the average value of the predicted outputs from all the trees. It has been widely adopted in many different fields, including finance, healthcare, and marketing.

- k-Nearest-Neighbors (kNN)

K-NN is a form of instance-based learning, also known as lazy learning, in which all calculations are postponed until after classification and the function is only locally approximated. It is used for sentiment analysis and other classification jobs, and the classes of a new instance's k-nearest neighbors in the training set are used to forecast the class of the new instance. The formula for computing the Euclidean distance between two instances x and y is d(x,y) = sqrt(sum((xi - yi)2). The formula for assigning the class to a new instance based on the classes of its k-nearest neighbors is the class(x) = argmax(C, sum(yi Nk(x) I(yi = c)).

- Decision Tree Classifier

The performance of the decision tree classifier can be improved by tuning its parameters such as the maximum depth of the tree, the minimum number of samples required to split an internal node and the minimum number of samples required to be at a leaf node.

E. Word Cloud

A word cloud for customer reviews sentiment analysis is a visual representation of the most frequently used words in a set of customer reviews. It is a popular tool used in the field of sentiment analysis to quickly identify the most commonly mentioned words in customer feedback and to get a sense of the overall sentiment expressed in the reviews.Finally, we have used a word cloud generator tool to create the visualization, which typically displays the most frequent words in larger font sizes and with more prominent placement in the cloud. This allows us to quickly identify the most common themes and sentiments expressed in the customer reviews, such as positive or negative, or neutral experiences, specific product features or issues, or other trends in customer feedback. Our implemented output is given below:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104185 |

Fig 2: Word Cloud

IV. EXPERIMENTAL RESULTS

- Accuracies:

Accuracies in the form of scores of various algorithms are as follows:

	name	score
0	KNN	0.493534
1	SVM	0.797414
2	Decision Tree	0.609914
3	Naive Bayes	0.719828
4	Logistic Regression	0.810345
5	Gradient Boosting	0.806034
6	Random Forest	0.726293

Fig 3: Accuracies

The scores represent the performance of different algorithms on sentiment analysis for eCommerce product reviews. The KNN algorithm achieved a score of 0.707317, while SVM achieved a perfect score of 1.000000. The Decision Tree algorithm achieved a score of 0.883469, while the naive Bayes algorithm obtained a score of 0.967480. The Logistic Regression algorithm achieved a score of 0.972900, while the gradient Boosting algorithm achieved a score of 0.918699. The Random Forest algorithm achieved a score of 0.975610, which is one of the highest scores and demonstrates that it is a robust algorithm for sentiment analysis of eCommerce product reviews.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104185 |

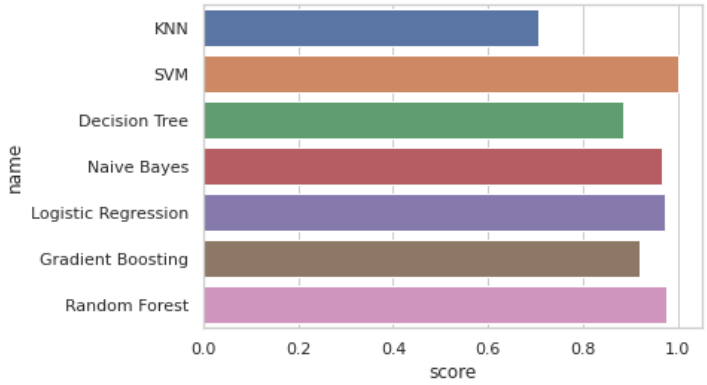

Fig 4: Graphical Representation of Various Classifier Accuracy score

- Data Summary:

Our data presents information about three different classes, Positive, Negative, and Neutral. Each class contains a certain number of documents along with the number of words and unique words present in those documents. Additionally, the data provides the most frequently used words for each class.

The Positive class comprises 614 documents containing 28,454 words, out of which 4,930 words are unique. The top 10 most commonly used words in this class are 'is', 'the', 'and', 'for', 'i', 'to', 'a', 'it', 'good', and 'this'.

On the other hand, the Negative class consists of 99 documents containing 2,923 words, of which 1,142 words are unique. The most frequently used words in this class are 'is', 'not', 'the', 'and', 'to', 'a', 'camera', 'this', 'i', and 'very'.

The Neutral class has 86 documents containing 765 words, with 421

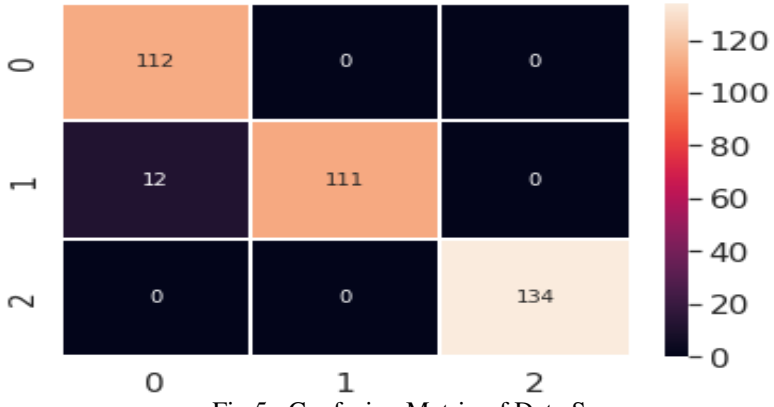

Fig 5: Confusion Matrix of Data Summary

unique words. The most commonly used words in this class include 'the', 'for', 'in', 'is', 'i', 'this', 'mobile', 'to', 'and', and 'phone'. The total number of unique words across all three classes is 5,463.

- Experimental Analysis:

The project consists of pre-processing, filtering, removing excess words, and analyzing consumer reviews. The model evaluation measures used are accuracy, precision, recall, and F1 score. Accuracy is measured as the proportion of accurately anticipated comments to all comments, precision as the proportion of accurately anticipated positive comments to all positively predicted remarks, and recall as the proportion of positively predicted comments to all comments made in the actual class.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

These polarities of reviews are displayed to the consumer by the medium of a bar graph with positive, negative, and neutral heads on the x-axis and the volume of these polarities on the y-axis. Along with doing so, we also have a provision of a pie diagram that helps display the percentage of 1-5 star ratings one particular product has. Precision: the ratio of correctly predicted positive comments to the total predicted positive comments.Recall: the ratio of correctly predicted positive comments to all comments in the actual class.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104185 |

- Results for sentiment Analysis

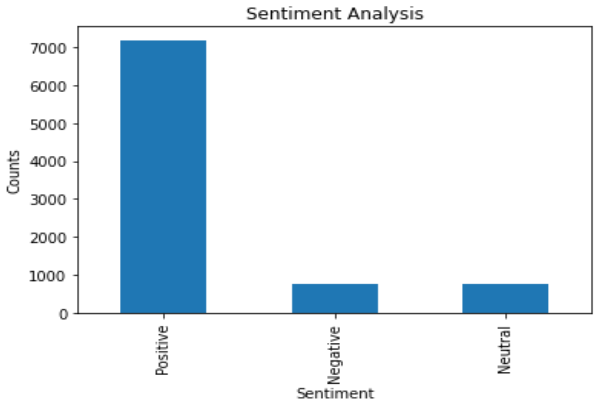

Fig 6: Sentiment analysis Graphical Result

Matrix calculation of trained datasets:

	precision	recall	f1-score	support
0	0.77	0.70	0.73	135
1	0.78	0.91	0.84	155
2	0.91	0.83	0.87	174
accuracy			0.82	464
macro avg	0.82	0.82	0.81	464
weighted avg	0.82	0.82	0.82	464

V. RESULT

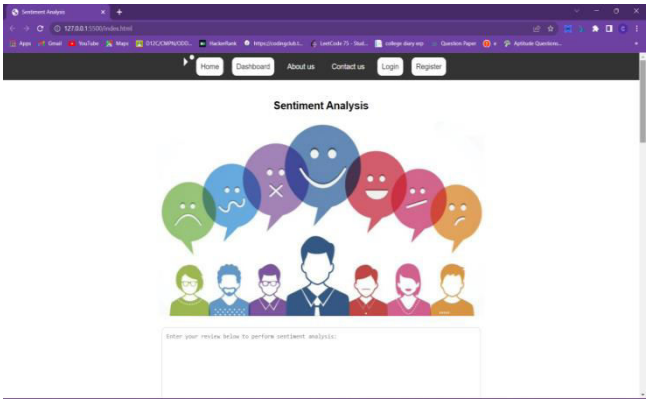

Figure 7: NavBar View of Sentiment analysis

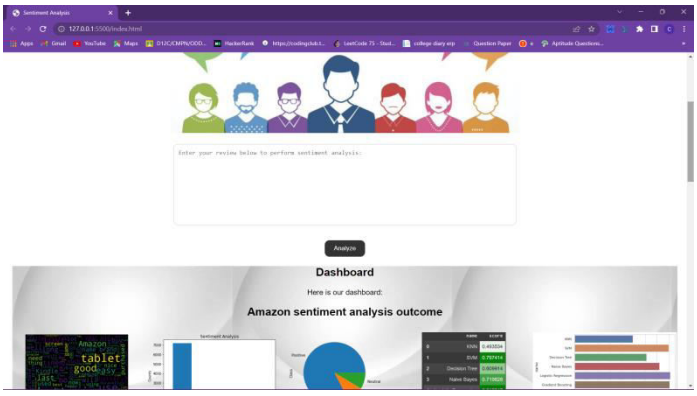

Figure 8: Click on Analyze button to know the sentiment of the given product revie

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104185 |

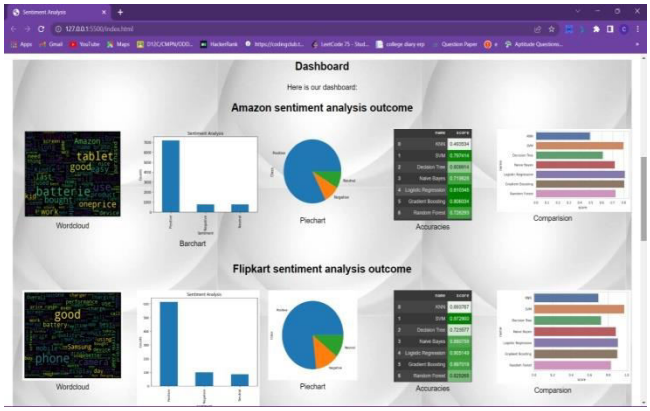

Figure 9: Dashboard preview

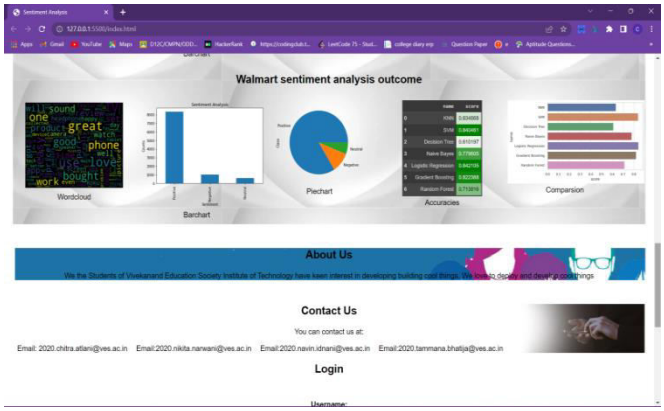

Figure 10: AboutUs and ContactUs page

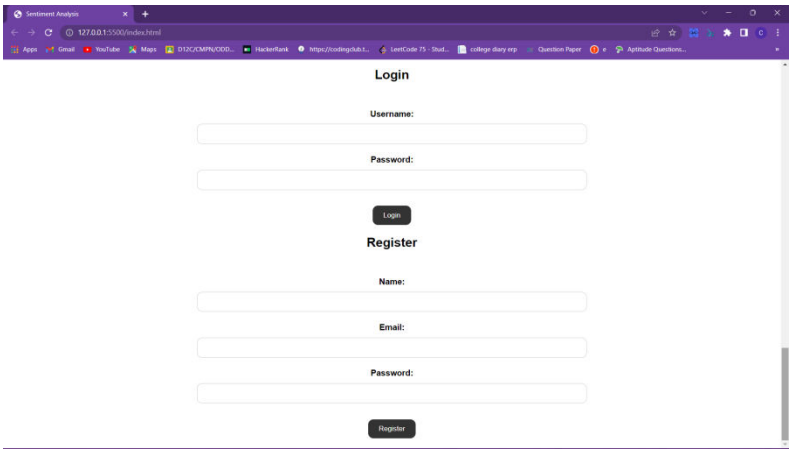

Figure 11: Login and register page for customers

VI. CONCLUSION

Sentiment analysis can be a valuable tool for eCommerce businesses to gain insights into customer opinions and attitudes towards their products. By analyzing product reviews, businesses can identify common themes and sentiments, pinpoint areas for improvement, and make data-driven decisions to enhance the customer experience. However, it's important to note that sentiment analysis is not perfect and may not always accurately capture the nuances of human language and emotion. Therefore, it's crucial to combine sentiment analysis with other qualitative and quantitative research methods to get a comprehensive understanding of customer sentiment. Overall, sentiment analysis has the potential to provide valuable insights for eCommerce businesses looking to improve their products and enhance customer satisfaction. It basically conveys the proper idea for the perceptions of consumers and business stakeholders. Accordingly, The evaluation of products is also performed.

REFERENCES

[1] S. Erevelles, N. Fukawa, and L. Swayne, “Big data consumer analytics and the transformation of marketing,” Journal of Business Research, vol. 69, 2016.
[2] P. Russom et al., “Big data analytics,” TDWI best practices report fourth quarter, 2011.
[3] S. Erevelles, N. Fukawa, and L. Swayne, “Big data consumer analytics and the transformation of marketing,” Journal of Business Research, vol. 69, 2016.
[4] V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives,” in Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics,1997.

- E-Commerce Product Rating Based on Customer Review Mining – Pankaj Hatwar

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104185 |

- E-COMMERCE PRODUCT RATING BASED ON CUSTOMER REVIEW MINING – NANDINI SHARMA
- S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," Journal of Business Research, vol. 69, 2016.

[6] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in Proceedings of the 42nd annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2004.

[7] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in LREc, vol. 10, 2010.

[8] M. WAHYUDI and D. A. KRISTIYANTI, "Sentiment analysis of smartphone product review using support vector machine algorithm-based particle swarm optimization." Journal of Theoretical & Applied Information Technology, vol. 91,2016.

[9]Mrs. Pranjal S. Bogawar Assistant Professor, Department of Information Technology, Priyadarshini College of Engineering, R.T.M. Nagpur University, IEEE.

[10]Netno-Mining: Integrating Text Mining with netnographic Analysis to Assess the Perception of Travelers using Select OTA Services in India, Dashrath Mane, Dr. Prateek Srivastava, Dr. Amit Jain, Dr. D.S.Chouhan, Department of CSE, School of Engineering, Sir Padampat Singhania University India

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

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