Study on Indian Buyers' Mindset in Context of Vocal for Local Using Sentiment Analysis

Beena Kapadia
Department of Computer Application, Career Point University, and IT Dept., VSIT Wadala,
Dr. Amita Jain
Information Technology Department, Vidyalankar School of Information Technology – Wadala,

Abstract
Atmanirbhar Bharat Abhiyan or Self-reliant India is the campaign initiated by the Hon'ble Prime Minister of India Shri Narendra Modi. Sentiment analysis is one of the Natural Language Processing Techniques, to understand and analyse the sentiment of the people which is used in finding the reviews of various products, places, or events. In this paper, the popularity of Indian Brands is studied using sentiment analysis techniques with respect to the trendy slogan 'Vocal for Local' given by our Hon'ble Prime Minister.The ratio used in the research is 70% data as training and 30% data will be testing. Two datasets have been used. One dataset consists of 696 respondents'sentiments collected across 49 cities of India for various products. To construct the new feature - sentiment score, VADER sentiment algorithm is usedto provide initial labelling of sentiment text. Also, some respondents have provided the Indian Brand name as the sentiment text. To tackle such responses,additional algorithm is combined along with VADER sentiment algorithm, and thus developed customised VADER sentiment algorithm.The other dataset contains text sentiments having 7215 rows of textto which customised VADER algorithm is applied. We achieved 88.75% accuracy on an average.
Keywords: 1. Atmanirbhar Bharat Abhiyan, 2.Self-reliant India, 3.Sentiment analysis, 4.Indian Brands, 5.VADER sentiment analysis, 6.Natural Language Processing.

Introduction

India and the entire world have suffered a lot due to the outbreak of pandemic and the resultant lockdown in 2020.[1] People suffered on health front as well as on economic front, which affected badly on the whole world economically resulting in fall of GDP of all major countries and a large-scale loss of human life in the world.

Our PM Narendra Modi came up with a call of 'Atmanirbhar Bharat', which means Self-reliant India or self-sufficient India. The basic idea behind this concept is making India self-generating economy. The short term goals of this program are economy, infrastructure, technology driven systems, vibrant demography and demand. [2]

Various slogans initiated under Atmanirbhar Bharat Abhiyan including 'Vocal for Local', 'Local for Global' and 'Make for World'. [2]. This paper focusses on finding out the popularity of Indian Brands and the mindset of Indians after pandemic in the view of Vocal for Local. Self-reliant program would help in reviving the sectors of Indian economy in short term goal of this program and in long term it will build capacities and make the country strong enough to face unprecedented situations like COVID. The program emphasizes on production locally.

In India, employment generation has always been a major concern for central and state government both, in public sector as well as in private sector. Self-reliant India can be a boon for young India. Besides encouraging entrepreneurship, the program will create jobs for all kinds of people – skilled or unskilled.

Cutting-edge technology like Artificial Intelligence (AI), Machine Learning (ML), Robotics, Deep Learning, Data Science, Cloud Computing have become important elements of planning, production, and services.

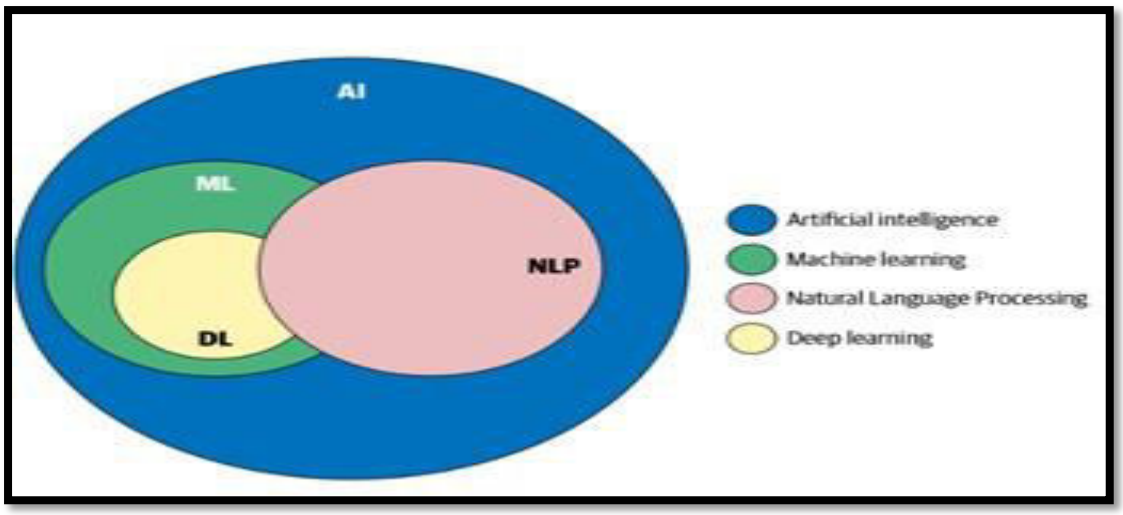


Figure 1: Relationship between AI, ML, DL and NLP[3]

artificial intelligence (AI) aims to build systems that can perform tasks which require human intelligence. [3]. Initial AI was mostly built out of logic, heuristics, and rule-based systems. Machine learning (ML) is a branch of AI that deals with the algorithms that can learn to perform tasks automatically based on huge amount ofdata, without requiring handcrafted rules. Deep learning (DL) is the branch of machine learning that is based on artificial neural network architectures. ML, DL, and NLP are all subfields within AI, and the relationship between them is depicted in Figure 1.

Natural Language Processing (NLP) is all about how a computer work with human languages. Spell check and auto correct, Auto-generated video transcription, Virtual assistants like Amazon's Alexa or Microsoft's Cortana windows 10, Autocomplete the sentence, Suggested articles or web series, detecting spam emails, Speech recognition, chatbot, sentiment analysis, bias in tweets to improving accessibility for people with disabilities are all examples of NLP.

Natural Language Processing is a branch of Artificial Intelligence that deals with the interaction between systems and humans using the natural language. The objective of Natural Language Processing is to read, decode, understand and make sense to perform desired task. The more data you collect, the more you can correct your algorithm's mistakes and reinforce its correct answers. [4]

NLP can be conducted in several programming languages. However, Python has some of the most extensive open-source NLP libraries, including the Natural Language Toolkit or NLTK. In this paper, Python is used to get all findings.

Text information can be extracted of from various sources like twitter, Facebook, Instagram, YouTube etc. in the form of comma separated values (.csv file) or tab separated values (.tsv file), for personal or commercial use. Sentiment Analysis (SA) classifies the sentiment of a people, which can be the positive sentiments, the negative sentiments, or the neutral sentiments. Many people tweet on twitter, which is the online source of data and can be used as one of the sources to analyse sentiments. This can be utilized to give various organizations an insight into their products, news channels can predict on polls, Hotels' review and many more such analysis can happen. Generally, SA is used to find the reviews of people. [4]

Now a days what data we get is not all in the clean text format. It can be from social media like Twitter, Facebook, Instagram or email data or some data we get from the website, which we may process online in real life applications, which is higher in volume, velocity and varieties, what we call as big data. We analyse to find the patterns from it and based on which we can make some useful predictions. In recent years, there has been a considerable rise of social media users and thus generating a massive amount of big data [5].

It is very important to pre-process the data before we perform sentiment analysis to it with any algorithm. There are various data Preprocessing techniques, to clean the data. If preprocessing is done properly to clean the data, the chances are quite high to get the maximum accuracy from that algorithm. Though various algorithms what we use to classify the data plays important role in finding the correct class, but data must be

cleaned and pre-processed to make it ready to get inputted in the required algorithm. There can be a huge difference in the results, if we use the data without preprocessing. [5]
In the next section we provide research objective, literature review, followed by the related work on primary data, Discussion and analysis and at the end Conclusion and Future Work.

Literature Review

A. Text Analysis Steps

A text analysis consists of three important steps: parsing, searching & retrieval, and text mining [6]. Parsing coverts unstructured text to a structured text for further analysis. Searching and retrieval are used to identify the text that contains search items. Search items can be specific words, phrases, or entities like product, people, or organizations. The of Part of Speech tagging is used to build a model, which takes input as a sentence, but the output is a tag will be sequence having various tags like Noun, Pronoun, Verb, Adverb etc. The POS tag assigns the appropriate part of speech for the corresponding word, according to the Penn Treebank POS tags. In this study, noun is very useful as various product names, or the word 'India' is used frequently. [7]. Stemming and Lemmatization are techniques to reduce different forms of the same word to the base form and that way to reduce the number of dimensions of the document. After Stemming and Lemmatization, raw text is transformed with text normalization techniques such as tokenization and case folding. Tokenization separates words from the body of text. Raw text is then converted into collections of tokens after performing tokenization step. Normally, each token is a word. Case folding reduces all letters to lowercase. Bag-of-words represents single words as identifiers. Using bag-of-words, the term frequency (TF) of each word can be calculated. Term frequency finds the weight of each term in a document. TF is proportional to the number of occurrences of the term in that document.[6]. Term Frequency-Inverse Document Frequency (TFIDF) directly works on top of the fetched documents. TFIDF treats these documents as the corpus. Sentiment analysis is performed on this pre-processed corpus.

B. Multinomial Naïve Bayes

Multinomial Naïve Bayes classifier works on the concept of term frequency (TF). TF means the number of times does the word occur in a document. Multinomial Naïve Bayes is used to find two facts that whether the word exists in a document or not as well as that words frequency in that document. [8].

C. Support Vector Machine

Support Vectors are simply the coordinates of individual observation. It is a classifier is a frontier which best segregates the classes.In SVM, it is required to select the hyper-plane which segregates the two classes better. [9]If we can't find linear hyperplane between the two classes, then it is required to introduce additional feature.

D. Random Forest

Random Forest algorithm is a supervised machine learning algorithm used for classification. It works based on the concept of ensemble learning, in which number of decision trees get various subsets of the data set. Here, all decision trees predict the output for the new data and the final class of that new data is considered as most of the outcome predicted from all decision trees.[10]

E. K Nearest Neighbor

The K Nearest Neighbor algorithm takes all the given data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a category, which is most suitable. That is, this algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much like the new data.
This algorithm first selects the number K of the neighbors (to get the exact k-value we need to test the model for each expected k-value.). Then, it calculates the minkowski distance of K number of neighbors. After that, it takes the k Nearest Neighbors as per the calculated minkowski distance. Among the identified k neighbors k neighbors, it counts the number of the data points in each category. It then assigns the new data points to that category for which the number of the neighbor is maximum.[11]

F. VADER Sentiment Analysis

VADER is abbreviation for "Valence Aware Dictionary and sEntiment Reasoner" and is available under the MIT License. The VADER tool was released in 2014. It uses a lexicon driven approach and additional heuristics for rating the input. It offers consistent ratings and requires no training data, as VADER is not a machine learning approach. It achieved some remarkable scores for multiple domains such as tweets, movie or product reviews. [12]

VADER is a rule-based sentiment analysis tool to express the sentiments on given text. [13] It is used to label the dataset into positive or negative sentiment score, based on whether its value exceeds 0.5 or not. The compound score can be calculated as the sum of all lexicon ratings which are normalized one. VADER works better than TextBlob for the text, taken from either social media or any web sources.[14]

G. Performance Evaluation

True positive means model has predicted positive and actual label is also positive. False positive means model has predicted positive but actual label is negative. True negative means model has predicted negative and actual value is also negative. False negative means model has predicted negative but actual label is positive. Accuracy is the ratio of the True predicted values to the Total predicted values. Accuracy = (True Positive + True Negative) / (True Positive + False Positive + True Negative + False Negative). Precision is how many values actually belong to the particular class out of all predicted class values. Precision = TP / (TP + FP) Recall is the ratio of TP to the actual value of all positive labels. That is Recall = TP / (TP + FN). F1-Score is calculated as the harmonic mean of precision and recall. It is 2 * (precision * recall)/ (precision + recall) [6].

Research Methodology:

A. Various Products

Generally, the products we use in day-to-day life are considered for the study. In this study some expensive Indian brands are considered like Allen Solly, Provogue, Biba, GINI & JONY, Van Heusen for clothes, Titan, Fastrack, Sonata for watches, Bajaj, TVS, Tata, Mahindra for automobiles, Tanishq, Kalyan Jewellers, Waman Hari pethe, Nakshatra Jewellers for Jewellery and Campus, Liberty, Woodland, Lakhani, Sparx for shoes.

There are some inexpensive Indian Brands also included in this study like Hamam, Santoor, Lifebuoy, khadi for soap, Khadi, Ayur, Patanjali, Indulekha, Dabur for shampoo / hair cleanser, Cello, Camlin, Flair, Natraj for pen / pencil, Nykaa, Biotique, Lakme, Elle18 for Lipstick and Organic India, Wagh Bakri Tea, Tata, Tulsi Tea for Tea or coffee.

Also, study has been conducted on other Indian products like Diya, Colours, Lights, handicrafts and clothes made by Hathkargha Udyog (handloom). Further study is carried out to find any change in respondent's buying behavior during and after pandemic- like buying Indian masks only or tried to buy as much Indian products as possible, which respondent had not thought about it ever before pandemic.

B. A typical sentiment analysis model:

A typical sentiment analysis model is designed as shown in the figure 2 below:

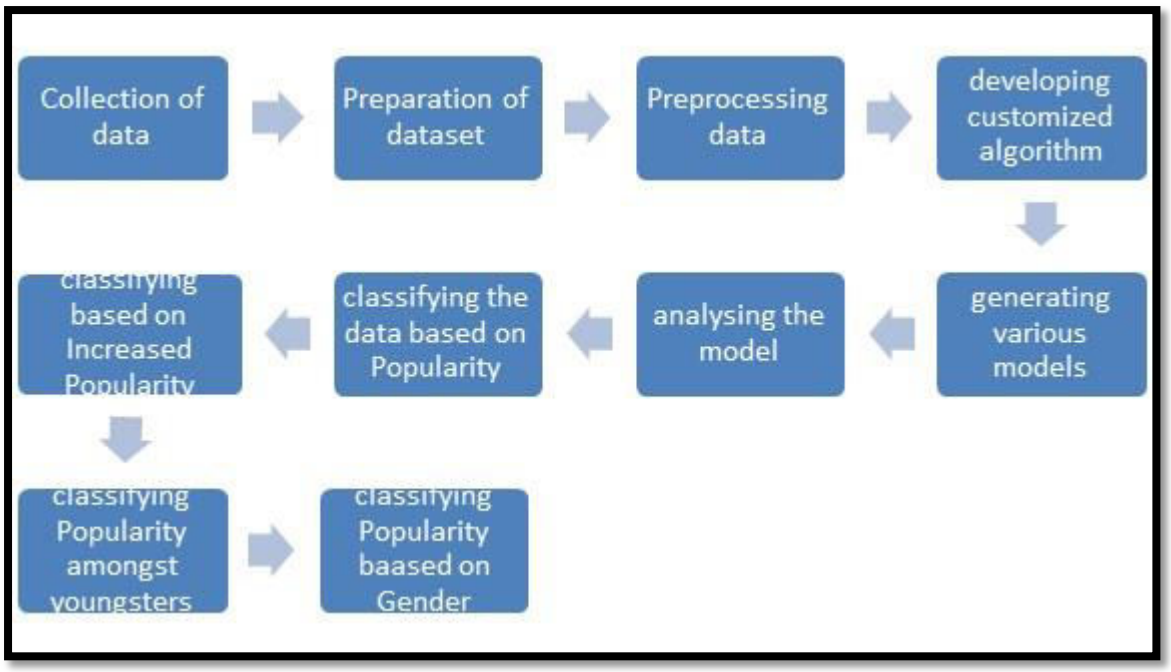


Figure 2: A typical sentiment analysis model

C. Collection of data

A questionnaire prepared and collected data from people all around India.The data set collected without any sentiment score information. It has been collected from total 49 cities of India. Total 696 data could be collected. It was filtered based on non-relevant data. So, after filtering total 656 data considered for the study. This collection of data is related to effectiveness and popularity of Atmanirbhar Bharat Abhiyan. So, data collection made considering the usage of various Indian Brands and general sentiment about any Indian brands.We collected a dataset of 696 data, but for this study data, we filtered the data to 656 as 40 respondents were unaware about 'Vocal for Local' or Atmanirbhar Bharat Abhiyan. After processing each ten products data, we merged all 656 data of all ten products totalling to 6560 data. Again, to those 6560 data, we merged the data, which was asked about responentds' sentiments of any Indian Brand. That way we could prepared a final dataset of total 7215 data.

D. Preparation of dataset

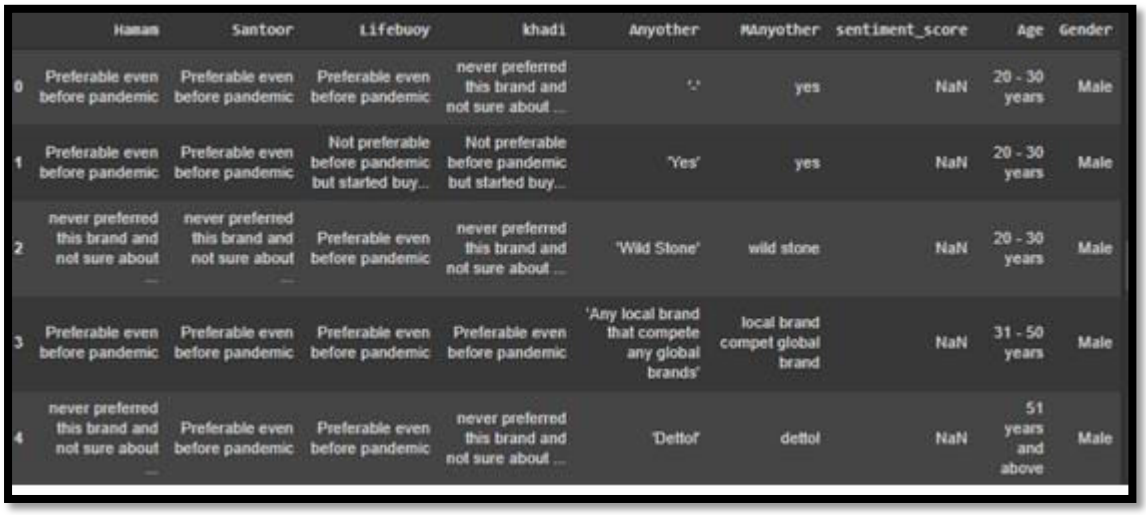
There are more than 80 columns in one data raw. So, processing all together is time consuming and complicated. So, data raw first divided Indian Brand wise. And their outcome is found. In the following figure, first 5 rows have been shown. The column headings are the name of different Indian soap brands used for this study, which are as Hamam, Santoor, Lifebuoy, and Khadi and column name 'Anyother' for respondent's choice of text. The Input data file is depicted in figure 3.

	Hamam	Santoor	Lifebuoy	khadi	Anyother	MAnyother	sentiment_score	Age	Gender
0	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	never preferred this brand and not sure about ...	-	-	NaN	20 - 30 years	Male
1	Preferable even before pandemic	Preferable even before pandemic	Not preferable before pandemic but started buy...	Not preferable before pandemic but started buy...	Yes	Yes	NaN	20 - 30 years	Male
2	never preferred this brand and not sure about ...	never preferred this brand and not sure about ...	Preferable even before pandemic	never preferred this brand and not sure about ...	Wild Stone	Wild Stone	NaN	20 - 30 years	Male
3	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	Any local brand that compete any global brands	Any local brand that compete any global brands	NaN	31 - 50 years	Male
4	never preferred this brand and not sure about ...	Preferable even before pandemic	Preferable even before pandemic	never preferred this brand and not sure about ...	Dettol	Dettol	NaN	51 years and above	Male

Figure 3: Input data file soap.csv for Indian Brand

E. Preprocessing data

Once a separate data file is prepared, it is used in our study. For each product, 4 to 5 Indian Brand choices were given to respondents, but the last column was for them to write their own view/ other brand name for that product. For example, for buying soap, the choices of Indian brands are given as Hamam, Santoor, Lifebuoy and khadi and the 5th column for soap is any other Indian Brand soap the respondent prefers. In 5th column respondent can write their expression, which is to be studied with sentiment analysis. The expression can be any other Indian soap brand name, which respondent prefers or any positive sentiment about any Indian soap brand or negative sentiment about any Indian soap brand. Various data preprocessing techniques applied to clean the text like replaced non alphabets to blank space, converted each sentence (review) split into different words (list), removed stop words from the word list, joined the words to frame sentence again and squeezed number of spaces into a single space.



	Hamam	Santoor	Lifebuoy	khadi	Anyother	MAnyother	sentiment_score	Age	Gender
0	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	never preferred this brand and not sure about ...	'-'	yes	NaN	20 - 30 years	Male
1	Preferable even before pandemic	Preferable even before pandemic	Not preferable before pandemic but started buy...	Not preferable before pandemic but started buy...	'Yes'	yes	NaN	20 - 30 years	Male
2	never preferred this brand and not sure about ...	never preferred this brand and not sure about ...	Preferable even before pandemic	never preferred this brand and not sure about ...	'Wild Stone'	wild stone	NaN	20 - 30 years	Male
3	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	'Any local brand that compete any global brands'	local brand compet global brand	NaN	31 - 50 years	Male
4	never preferred this brand and not sure about ...	Preferable even before pandemic	Preferable even before pandemic	never preferred this brand and not sure about ...	'Dettol'	dettol	NaN	51 years and above	Male

Figure 4: column 'Anyother' and column 'MAnyother' – which is modified column 'Anyother' to perform sentiment analysis.

The MAnyother is modified Anyother column, in which '-', blank or any punctuation is replaced with either yes or no based on the choices selected by respondent from column 0,1,2 and 3 – the brands of soaps; which is shown in figure 4.

F. Developing customized VADER Sentiment algorithm to construct a feature Sentiment Score

While collecting the data, the sentiment score was not asked to respondents as 1, or -1 or zero. The challenge was to develop the logic to construct the correct sentiment score.

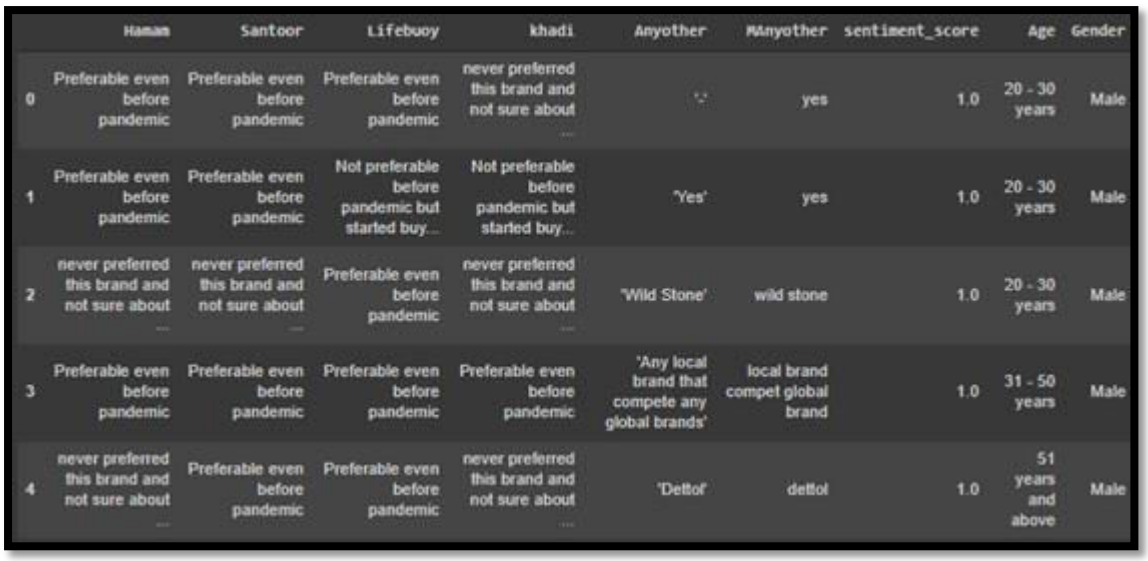
When applied TextBlob library to the 5th column, which is the text entered by respondents, to find sentiment score, we didn't get the correct sentiment score for all data rows. It was not giving sentiment even for the words 'yes' or 'no'. There must be happy, awesome kinds of words for positive sentiment and bad, worst kind of words for negative sentiment. By applying VADER Sentiment Algorithm to text data, we could get correct sentiment for the text in which clearly, 'yes' or 'no' were mentioned.

The output of applying VADER sentiment algorithm, without customized VADER sentiment analysis is shown in figure 5. Where in 0th row, the soap brand Hamam, Santoor and Lifebuoy are preferable by respondent, hence, its sentiment should be positive; even though in Anyother column '-' is written. Here, we have applied the condition, which states that if respondent uses any of the four given Indian Brands, then '-' or blank or any punctuation should be replaced with yes; and if all four not selected then the punctuation or blank space should be replaced with 'no'. In rows 2 and 4, the other Indian Brand name is mentioned - 'Wild Stone' and 'Dettol' respectively, so the respondent uses Indian brand only, but the other Indian brand. So, ideally sentiment should be positive. But as the brand name doesn't show any sentiment in general, so, the sentiment score using VADER sentiment analysis is neutral. Also, in row 3, respondent has conveyed 'Any other local brand that compete any global brands', also has positive sentiment towards buying Indian soap brand. But the sentiment score for the same is 0.0, that is neutral.

	Hamam	Santoor	Lifebuoy	khadi	Anyother	MAnyother	sentiment_score	Age	Gender
0	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	never preferred this brand and not sure about ...	'-'	yes	1.0	20 - 30 years	Male
1	Preferable even before pandemic	Preferable even before pandemic	Not preferable before pandemic but started buy...	Not preferable before pandemic but started buy...	'Yes'	yes	1.0	20 - 30 years	Male
2	never preferred this brand and not sure about ...	never preferred this brand and not sure about ...	Preferable even before pandemic	never preferred this brand and not sure about ...	'Wild Stone'	wild stone	0.0	20 - 30 years	Male
3	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	'Any local brand that compete any global brands'	local brand compet global brand	0.0	31 - 50 years	Male
4	never preferred this brand and not sure about ...	Preferable even before pandemic	Preferable even before pandemic	never preferred this brand and not sure about ...	'Dettol'	dettol	0.0	51 years and above	Male

Figure 5: Applied VADER Sentiment algorithm to generate sentiment score

To overcome that problem and to get the correct sentiment score of the respondent, we used the concept of 'Named Entity Recognition'[7]. We can get the correct sentiment based on the data by using customized VADER Sentiment Analysis. There were two situations, where VADER sentiment algorithm didn't work properly. First is when some noun is given. The noun is nothing but the Indian product name. By writing Indian Product / Brand name, the respondent supports Indian brand only, but it doesn't have happy / sad or yes/no kind of answer. By applying the logic of 'if content is a noun, then change the sentiment score to 1' to where it was zero sentiment by VADER sentiment algorithm, ultimately correct emotions can be retrieved. Secondly, if respondent has already selected any Indian Brand from the given list then he / she may write '-' or 'no' for one more brand in the text. So, we considered previous choices also, if sentiment score is resulted to zero by VADER sentiment algorithm. Thus, we developed a customized VADER algorithm for our study as shown in figure 6. The logic of customized VADER algorithm is shown in figure 7.



	Hamam	Santoor	Lifebuoy	khadi	Anyother	MAnyother	sentiment_score	Age	Gender
0	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	never preferred this brand and not sure about ...	'-'	yes	1.0	20 - 30 years	Male
1	Preferable even before pandemic	Preferable even before pandemic	Not preferable before pandemic but started buy...	Not preferable before pandemic but started buy...	'Yes'	yes	1.0	20 - 30 years	Male
2	never preferred this brand and not sure about ...	never preferred this brand and not sure about ...	Preferable even before pandemic	never preferred this brand and not sure about ...	'Wild Stone'	wild stone	1.0	20 - 30 years	Male
3	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	Preferable even before pandemic	'Any local brand that compete any global brands'	local brand compet global brand	1.0	31 - 50 years	Male
4	never preferred this brand and not sure about ...	Preferable even before pandemic	Preferable even before pandemic	never preferred this brand and not sure about ...	'Dettol'	dettol	1.0	51 years and above	Male

Figure 6: Applied Customized VADER Sentiment algorithm to generate sentiment score

Using customized VADER Algorithm, we could generate the new column of sentiment score to work with classification model.

G. Generating various models

The processed text and sentiment score generated by customized VADER sentiment algorithm are applied to various models. Multinomial Naïve Bayes classification, Support Vector Machine algorithm[4], Random Forest algorithm and K Nearest Neighbor algorithm are used to evaluate the model.

Analysis of Models

Analysing the model for the first objective, which is to find out the popularity of Indian products due to "Atmanirbhar Bharat Abhiyan". The popularity of Indian Brands for Clothing, Watches, Automobiles, Jewellery and shoes are respectively 90.2, 98.8, 95.4, 90.2 and 94.2 respectively which is shown in the following table 1 and table 2 for expensive products and inexpensive products respectively. This data is taken between Feb 2022 to May 2022, which is

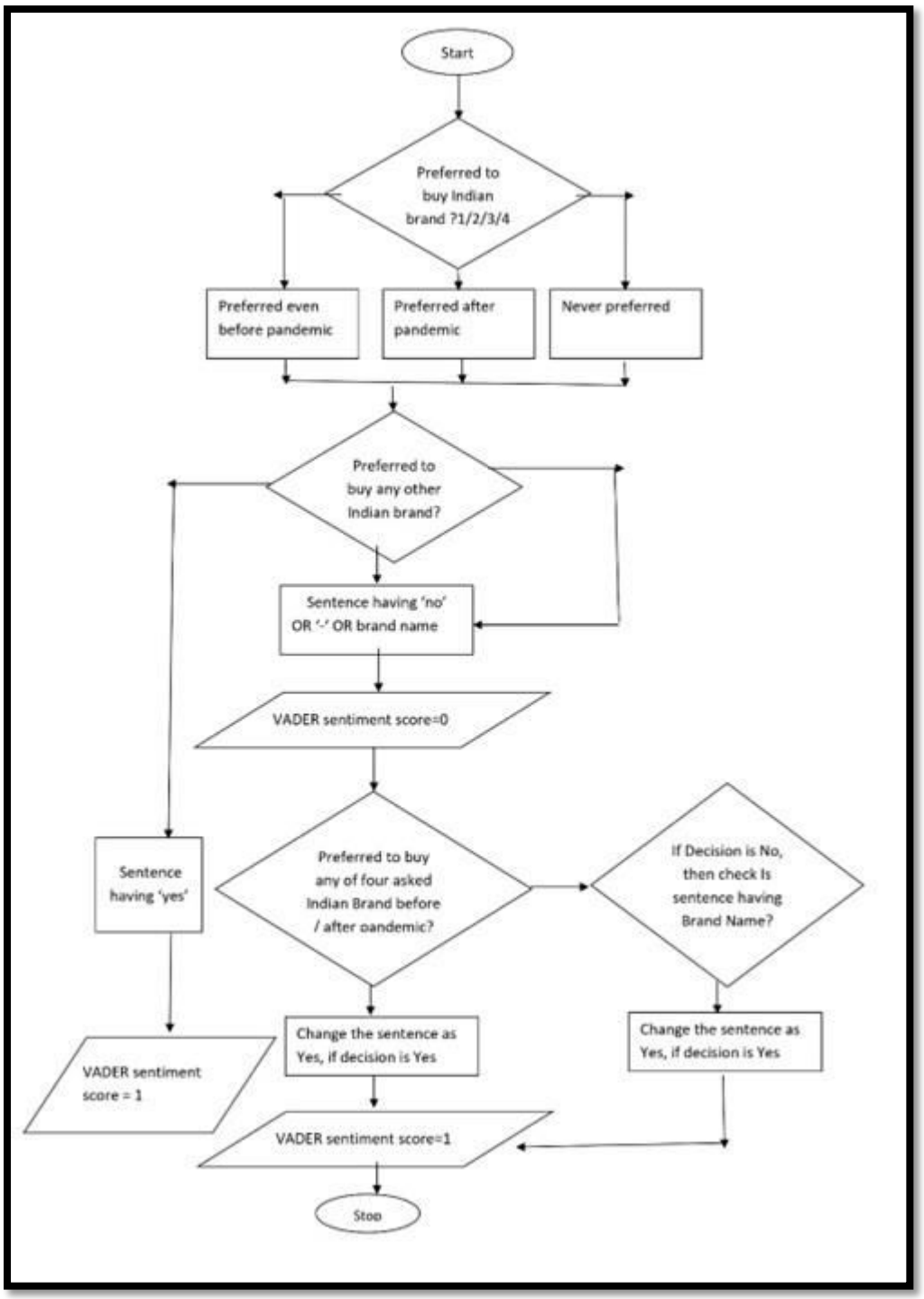


Figure 7: Flow chart of customized VADER sentiment analysis

after pandemic and a call from Hon'ble Prime Minister of India Shri Narendra Modiji. Moreover, in table 1 and table 2, age wise and gender wise classification also made.

Table 1: popularity of Expensive Indian Brands in India

Model accuracy for various Indian Brands / Model Name	Clothing	watches	Automobiles	Jewellery	shoes
Multinomial Naïve Bayes	0.90	0.99	0.95	0.90	0.94
SVM	0.90	0.99	0.95	0.90	0.94
RF	0.90	0.99	0.95	0.90	0.94
KNN	0.90	0.99	0.95	0.90	0.94
Overall Popularity of Indian Brands	90.2%	98.8%	95.4 %	90.2%	94.2 %
Below 20	94.4 %	94.4 %	88.9 %	88.9 %	88.9 %
Between 21 and 30	83.5 %	98.4 %	96.1 %	89.0 %	89.8 %
Between 31 and 50	92.6 %	100.0 %	94.6 %	90.6 %	91.9 %
51 and above	91.2 %	97.1 %	100.0 %	94.1 %	100.0 %
Male	76.9 %	76.3 %	78.1 %	78.1 %	75.1 %
Female	79.9 %	68.6 %	69.2 %	69.8 %	67.3 %

Table 2: popularity of Inexpensive Indian Brands in India

Model accuracy for various Indian Brands / Model Name	pen / pencil	Lipstick	Tea / coffee	soap or hand wash or face wash	shampoo / hair cleanser
Multinomial Naïve Bayes	0.98	0.83	0.95	0.96	0.84
SVM	0.98	0.84	0.95	0.96	0.85
RF	0.98	0.85	0.95	0.96	0.85
KNN	0.98	0.84	0.95	0.96	0.74
Overall Popularity of Indian Brands	98.8%	83.8%	94.8 %	96.4 %	85.7 %
Below 20	100.0 %	83.3 %	94.4 %	94.4 %	83.3 %
Between 21 and 30	99.2 %	77.2 %	96.9 %	96.1 %	82.7 %
Between 31 and 50	98.0 %	89.3 %	93.3 %	96.0 %	85.9 %
51 and above	100.0 %	85.3 %	94.1 %	100.0 %	88.2 %
Male	78.1 %	57.4 %	79.3%	79.9 %	66.9 %
Female	69.8 %	64.8 %	72.3%	74.8 %	65.4 %

The model accuracy is quiet good and almost in all cases any model reflected the same accuracy, which is quite satisfactory.

Further, working for the second objective to find whether this much popularity was there in India for Indian Brands before pandemic, or it increased after pandemic, is shown in the following table 3 and Table 4 for

Expensive items and inexpensive items respectively. Also, the popularity of Indian brands increased in the younger generation also – whose age is below 20, after pandemic. Also, men have started buying / liking more Indian brands after pandemic compared to women. There is a rise in the popularity of buying Indian products in general in everyone, whatever the gender may be or whatever the age may be.

Table 3: Popularity percentage of Expensive Indian Brands increased after pandemic

ensive Indian Brands / Not preferred before but started preferring after pandemic	Clothing	watches	Automobiles	Jewellery	shoes
Popularity percentage increased after pandemic	21.6 %	15.2 %	14.3 %	20.1 %	16.2 %
Popularity percentage increased after pandemic having age below 20	0.9 %	0.6 %	0.3 %	1.2 %	0.6 %
Popularity percentage increased after pandemic having age equal or above 20	20.7 %	14.6 %	14.0 %	18.9 %	15.6 %
Popularity percentage increased after pandemic in male	13.1 %	10.9 %	10.6 %	13.4 %	10.1 %
Popularity percentage increased after pandemic in female	8.5 %	4.3 %	3.7 %	6.7 %	6.1 %

Table 4: Popularity percentage of Inexpensive Indian Brands increased after pandemic

Expensive Indian Brands / Not preferred before but started preferring after pandemic	pen / pencil	Lipstick	Tea / coffee	soap or hand wash or face wash	shampoo / hair cleanser
Popularity percentage increased after pandemic	14.0 %	11.3 %	15.5 %	18.0 %	17.7%
Popularity percentage increased after pandemic having age below 20	0.6 %	0.3 %	0.9 %	0.6 %	0.6 %
Popularity percentage increased after pandemic having age equal or above 20	13.4 %	11.0 %	14.6 %	17.4 %	17.1 %
Popularity percentage increased after pandemic in male	10.0 %	7.0 %	10.6 %	12.5 %	11.3 %
Popularity percentage increased after pandemic in female	4.0 %	4.3 %	4.9 %	5.5 %	6.4 %

Discussionon analysis

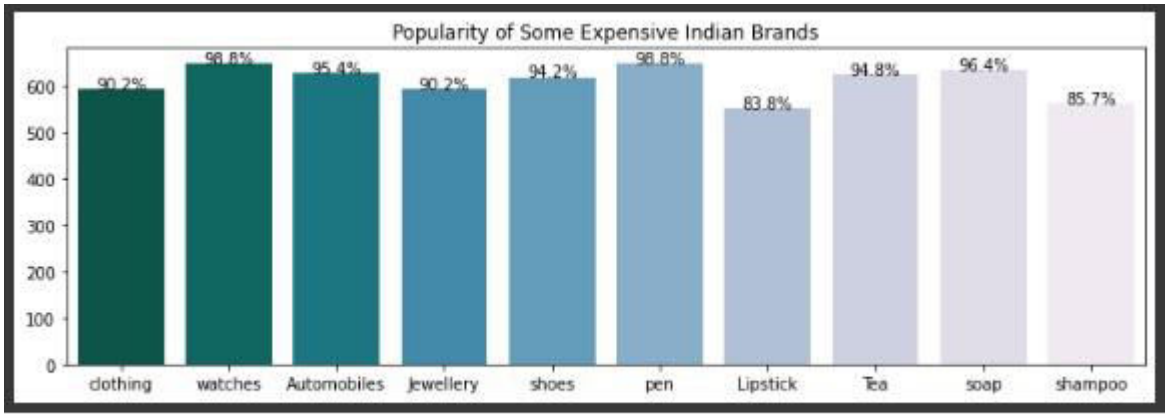


Figure 8: Popularity of Some Expensive Indian Brands

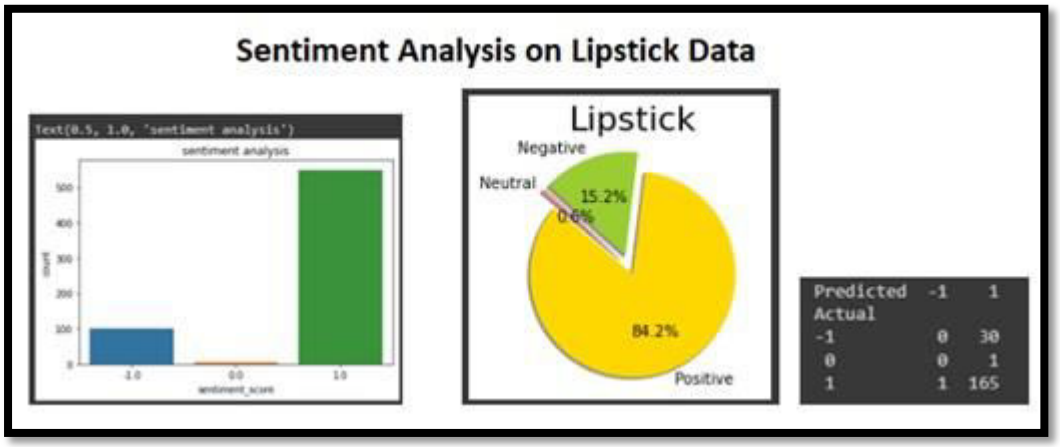


Figure 9: (a) Bar chart, (b) Pie chart and (c) Confusion matrix

Popularity of Indian Brands are high irrespective to whether it is expansive brand or inexpensive brand, which is shown in figure 8. Figure 9(a), (b) and (c) depicts the sentiment analysis on a particular product 'Lipstick' in the form of Bar chart Pie chart and confusion matrix respectively. There are 165 TP and 31 FP data: whereas 1 FN data out of total 197 test data (30%). Figure 10is the visualization of group bar chart for age wise popularity of Indian Brands after pandemic.

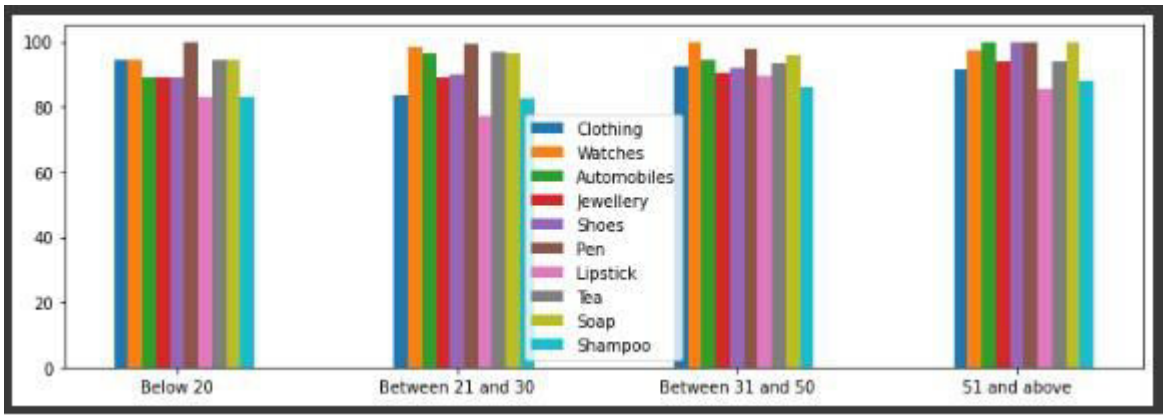


Figure 10: age wise popularity of Indian Brands

It is obvious from figure 11, Male prefers Indian brands for watches, Automobiles, shoes, Jewellery, pen, Tea, soap and shampoo are all popular amongst Men than amongst Female. Whereas female prefers Indian Brand Clothing and Lipstick more than Male.

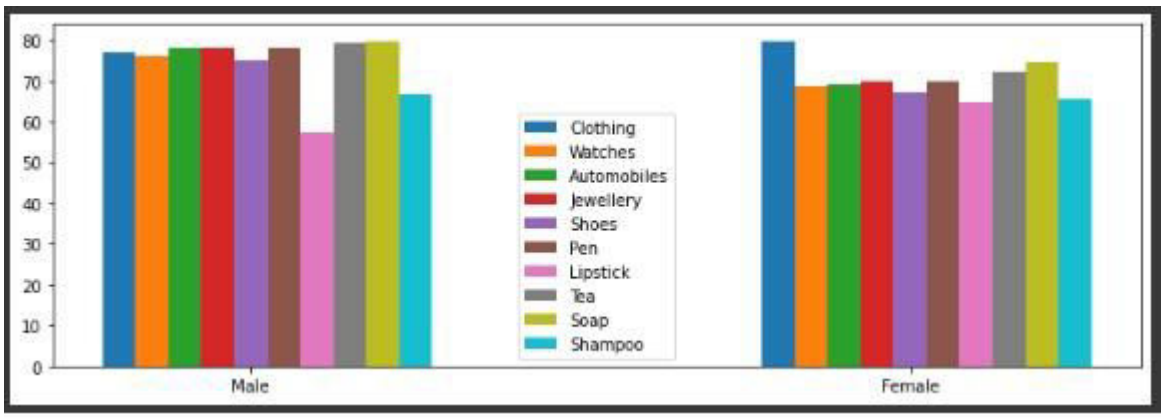


Figure 11: gender wise popularity of Indian Brands

When we compare exceed in popularity whether it is expensive or inexpensive brand as shown in figure 12, then in any Indian product, the popularity is certainly increased after a call of 'Vocal for Local' given by our Hon'ble Prime Minister during pandemic in Context of Atmanirbhar Bharat Abhiyan as clearly visible in figure 11, with the help of Matplotlib stacked bar chart with values on bar.

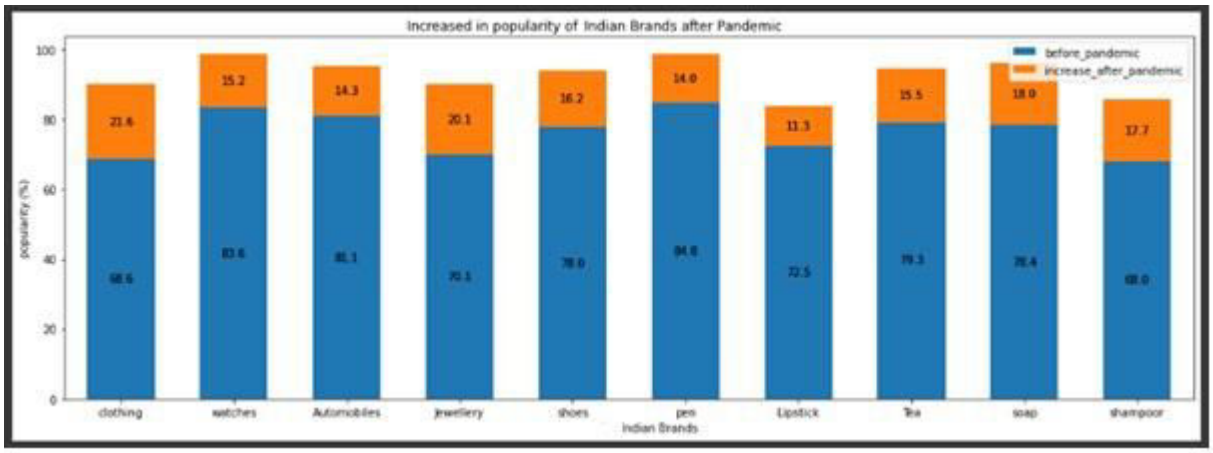


Figure 12: Exceed in popularity of Some Indian Brands, after 'Vocal for Local'

Figure 13 and Figure 14 shows how much increment in the popularity of Indian brands amongst youngsters and comparison between men and women in popularity of Indian Brands.

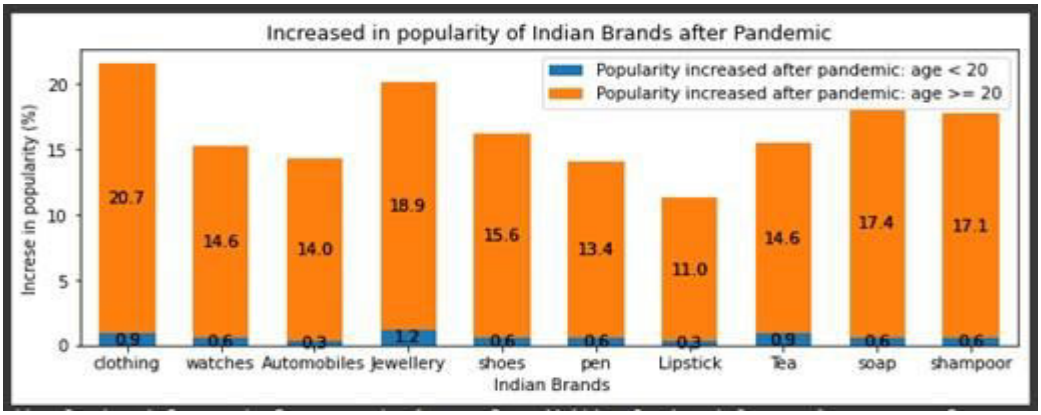


Figure:13 Increase in popularity of Indian Brands amongst youngsters compare to matured people

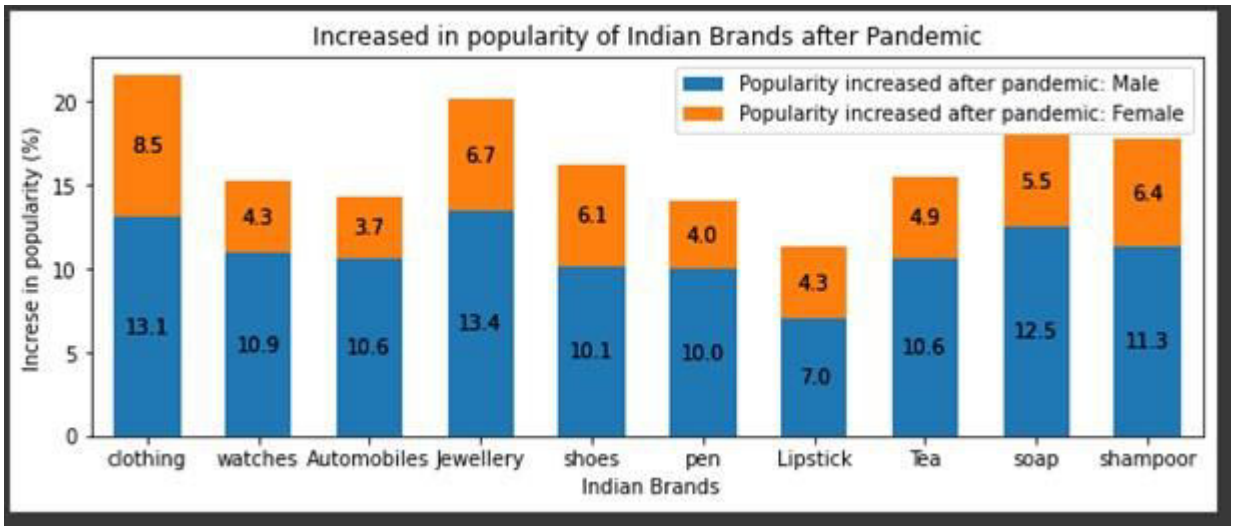


Figure:14 Increase in popularity of Indian Brands amongst Female and Male

By applying only VADER Sentiment analysis on the Text for the question 'Do you find any change in your buying behavior during and after pandemic- like buying Indian masks only or tried to buy as much Indian products as possible, which you had not thought about it ever before pandemic. Please share your experience.', the opinion what we received is as 24.7% positive, 21.6% negative and 53.7% neutral; with model accuracies are 73%,66%, 79%, 62% with SVM, Multinomial Naïve Bayes, Random Forest Classifier, and KNN respectively. But by applying customized VADER sentiment analysis we received is as 43.3% positive, 21.6% negative and 35.1% neutral; with model accuracies are 79%, 53%, 82%, 66% with SVM, Multinomial Naïve Bayes, Random Forest Classifier, and KNN respectively.
By taking a common dataset of 7215 data, when merged in one file for all 10 different products including general opinion about Indian Brands, we received output as shown in table 5 and figure 15.

Table 5: popularity of All 10 Indian products + any Indian product in general in India

el Name Model accuracy for various Indian Brands	MNB	SVM	RF	KNN
All 10 Indian products + any Indian product in general	87%	89%	90%	89%

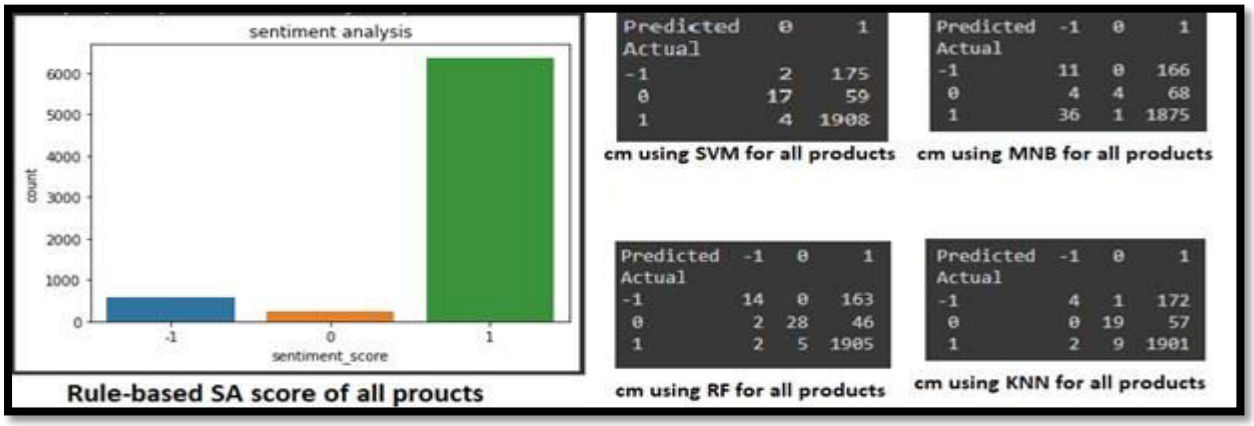


Figure 15: rule-based sentiment analysis score and confusion matrix (cm) using SVM, MNB, RF and KNN machine learning algorithms for respondents 'opinion on all 10 specific products the opinion on general Indian Brands

Conclusion and future work

By using TextBlob for finding sentiment score, on an average for all 10 products, we could get the average accuracy 60.3%and popularity as 56.88% only for any choice of the Indian products. Though it was not matching looking at the actual data. By applying VADER Sentiment algorithm instead of TextBlob, the average accuracy was 88.5 and the popularity increased by 9.8 %. But with customized VADER Sentiment Algorithm, the average accuracy could be achieved as 92.1 and the average popularity is 92.83 %. 67.5 percentage respondents believe that Vocal for Local is new form of globalization and not a rejection of globalization. 63.5% respondents believe that during COVID lockdown period Local Brands served better than Global Brands. 89.1 % respondents agreed about the fact that they bought Indian products like Diya, Colours, Lights etc. to celebrate Diwali last year. 69.3 % respondents feel that the Indian products and handicrafts taking over the market during festivals. 62.9 % respondents use clothes made by Hathkargha Udyog (handloom), out of which 8.3 % respondents were not preferring to use before pandemic but started buying after pandemic. More data can be collected for better outcome. Also, popularity of Indian Brands is more inclined towards men than women and though too much of popularity for Indian brands is not there amongst youngsters but at least some of them started liking Indian brands.

We achieved 89% accuracy using SVM, 87% accuracy using Multinomial Naïve Bays, 90% accuracy with RF, 89% accuracy with KNN on all 11 products data set of total 7215 data with rule-based popularity as 88.3% positive response, 8.2 % negative response and 3.5% neutral response and with Machine Learning techniques

References

1. *Panneer, S.; Kantamaneni, K.; Akkayasamy, V.S.; Susairaj, A.X.; Panda, P.K.; Acharya, S.S.; Rice, L.; Liyanage, C.; Pushparaj, R.R.B. The Great Lockdown in the Wake of COVID-19 and Its Implications: Lessons for Low and Middle-Income Countries. Int. J. Environ. Res. Public Health 2022, 19, 610.*
2. *V. Bonta, N. K. N. J. for Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis, Asian Journal of Computer Science and Technology vol. 8 pp. 1 -6, 2019.*
3. *Kapadia, B., Jain, A. (2021). Analysis of Papers Based on Sentiment Analysis Applications on E-Commerce Data. In: Abraham, A., Sasaki, H., Rios, R., Gandhi, N., Singh, U., Ma, K. (eds) Innovations in Bio-Inspired Computing and Applications. IBICA 2020. Advances in Intelligent Systems and Computing, vol 1372. Springer, Cham.*
4. *Pradha, S., Halgamuge, M. N., & Vinh, N. T. (2017). Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data. 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (p. 8). Da Nang, Vietnam: IEEE.*
5. *Data Science & Big Data Analytics Discovering, Analyzing, Visualizing and Presenting Data EMC Education Services by David Dietrich, Barry Heller, and Beibei Yang, Published by John Wiley & Sons, Inc. 10475 Crosspoint Boulevard Indianapolis, ISBN: 978-1-118-87613-8.*
6. *Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. 2019 International Conference on Automation, Computational and Technology Management (ICACTM).*
7. *Borg, A., & Boldt, M. (2020). Using VADER Sentiment and SVM for Predicting Customer Response Sentiment. Expert Systems with Applications, 113746.*
8. *Singh, S. N., &Sarraf, T. (2020). Sentiment Analysis of a Product based on User Reviews using Random Forests Algorithm. 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence).*
9. *[Irawaty, I., Andreswari, R., &Pramesti, D. (2020). Development of Youtube Sentiment Analysis Application using K-Nearest Neighbors (Nokia Case Study). 2020 3rd International Conference on Information and Communications Technology (ICOIACT).*
10. *Hutto, C.J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of social media Text. ICWSM.*
11. *Yang, S. E. (n.d.). Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment. In M. 1.-1. IMECS 2019 (Ed.). ISBN: 978-988-14048-5-5*