ISSN (online):2396-9889

INTERNATIONAL RESEARCH JOURNAL OF TECHNOLOGY AND APPLIED SCIENCE

http://www.irjtas.com

(An ISO Certified Journal)

VOL. 05 Issue 02 FEBRUARY 2021

Sentiment Analysis of Amazon Mobile Reviews Based on Feature Extraction Approach over Cloud Environment

Kapil Jain
MTech Scholar
kpljain21@gmail.com

Prof. Amit Ganguli
M. Tech. Co-ordinator
amitganguli@sistec.ac.in

Prof. Ajit Kumar Shrivastava
Head of Dept, CSE
SISTecR.hodcs@sistec.ac.in

Abstract*— Online shopping has been growing for 20 years and many e-commerce websites such as Amazon, have been created to meet the increasing demand. Consequently, a specific product can be bought on several websites and the prices may vary. As customers usually want the best quality for the lowest price but can't directly check it, reviews from other customers seem to be the most reliable way to decide whether to buy the product or not. Therefore, sentiment analysis has proven essential to understand a product's popularity among the buyers all over the world. Sentiment analysis is a classification process whereby machine learning techniques are applied on text-driven datasets in order to analyze its sentiment. But the data of ecommerce are growing rapidly and need high end processor to process these data. Distributed computing resembles a panacea to defeat the obstacles. It vows to expand the speed with which the applications are conveyed, expanded imagination, development, brings down cost at the same time expanding business sharpness. It calls for fewer ventures and a collect of advantages. The end-clients just compensation for the measure of assets they utilize and can without much of a stretch scale up as their necessities develop. Specialist co-ops, then again, can use virtualization innovation to expand equipment use and work on administration. In this, we will take an virtual machine on cloud preferred over others because cloud service is provided by third party providers (Google cloud platform (GCP)) so for security reason private cloud give better security than others because the connection between user and virtual machine is secured by ssh. And on private cloud we will easily scale the storage and processing power at any time whenever application required. and that we can build logistic regression model supported different feature extraction techniques like BOW (Bag of Words), TF-IDF and N-Gram , From experimental result we will say that model repose on N-Gram features provides better accuracy as compared to others.

Keywords: cloud computing, private cloud, machine learning application, security, ssh, feature extraction, Bag of Words, TFID, N-Gram.

I.INTRODUCTION

These days, a gigantic measure of surveys is accessible on the web. Besides offering a valuable source of information, these informational contents generated by users, also called User Generated Contents (UGC) strongly impact the purchase decision of customers. As a matter of fact, a recent survey revealed that 67.7% of consumers are effectively influenced by online reviews when making their purchase decisions. More precisely, 54.7% recognized that these reviews were either fairly, very or absolutely important in their purchase decision making. Depending on online audits has hence become a natural for customers.

In their examination interaction, customers need to discover helpful data as fast as could really be expected. However, searching and comparing text reviews can be frustrating for users as they feel submerged with information. Indeed, the massive amount of text reviews as well as its unstructured text format prevent the user from choosing a product with ease. The star-rating, i.e. stars from 1 to 5 on Amazon, rather than its text content gives a quick overview of the product quality. This numerical information is the number one factor used in an early phase by consumers to compare products before making their purchase decision.

However, many product reviews (from other platforms than Amazon) are not accompanied by a scale rating system, consisting only of a textual evaluation. In this case, it becomes daunting and time-consuming to compare different products in order to eventually make a choice between them. Hence, models ready to anticipate the client rating from the content survey are fundamentally significant. Getting a general feeling of a printed survey could thusly improve customer experience.

ISSN (online):2396-9889
INTERNATIONAL RESEARCH JOURNAL OF TECHNOLOGY AND APPLIED SCIENCE
http://www.irjtas.com
(An ISO Certified Journal)
VOL. 05 Issue 02 FEBRUARY 2021

Cloud Computing

Cloud Computing can be defined as the novel style of computing where virtualized resources are provided as services on internet which are dynamically scalable[1].cloud computing represents a different way to architect and remotely managing computing resources. It refers to both application delivered as the service over the internet and system software in the datacenters that provide those services .the data centre hardware and software is called cloud[2]. Cloud Computing is a major paradigm shift [3]. Most of the enterprises shifting their applications on to the cloud owing to its speed of implementation and deployment, improved customer experience, scalability, and cost control. Reliability, availability and security are the three greatest concerns for moving on to the cloud [3]. Businesses are running all kinds of applications in the cloud, like customer relationship management (CRM), HR, accounting, and much more. A portion of the world's biggest organizations moved their applications to the cloud with salesforce.com after thoroughly testing the security and unwavering quality of framework. Smart phones, laptops, PCS and PDAs can access programs, storage and application development platforms over the internet using cloud computing via services offered by the cloud providers. Virtualization is the key innovation that empowers Cloud Computing [3]. Far off facilitating took its change from leasing framework to giving and keeping up Virtual workers supporting the variances popular. The large parts in distributed computing are Google, Amazon, and, of late, Microsoft and IBM. The early adopter of this innovation is Amazon. Amazon started giving Amazon Web Services in 2005, known uniquely to the cognoscenti. Amazon's Web Services is the most established and generally develop of the public cloud specialist co-ops. Microsoft Azure addresses a significant development both of working frameworks and of Microsoft's general methodology. While composed totally starting from the earliest stage, it profits by a long, for the most part recognized, and costly family. Google was an early advocate of both virtualization and distributed computing.

DEPLOYMENT MODEL ON CLOUD

The deploy cloud computing in several different ways depending upon many factors, such as:

- Where the cloud services are hosted
- Security requirements
- Desire to share cloud services
- The ability to manage some or all of the services
- Customization capabilities

There are four basic sending models for cloud administrations

1. Public Cloud
2. Private Cloud
3. Community Cloud and
4. Hybrid Cloud

Public Cloud

Available to the general public and owned by a third party cloud service provider (CSP). The computing resources over the internet from a CSP, who shares it resources with other organizations. This is most cost effective deployment model. All administrations are conveyed with steady accessibility, strength, security and sensibility. The benefits of public cloud reduce and control monitoring over the provider's governance and security.

Private Cloud

It includes an unmistakable and secure cloud based climate in which just the predefined customer can work. However, this type of cloud is only accessible by a single organization providing that organization with greater control and privacy. The highlights and advantages of the private distributed computing are high security and protection, more control, cost and energy productivity, improved unwavering quality and cloud blasting.

Community Cloud

This processing is a cooperative exertion wherein foundation is divided among a few associations from a particular local area with normal concerns, regardless of whether oversaw inside or by an outsider and facilitated inside. The costs are spread over fewer users

ISSN (online):2396-9889
INTERNATIONAL RESEARCH JOURNAL OF TECHNOLOGY AND APPLIED SCIENCE
http://www.irjtas.com
(An ISO Certified Journal)
VOL. 05 Issue 02 FEBRUARY 2021

than a public cloud, so only some of the cost saving potential of cloud computing are realized.

Hybrid Cloud

It is an integrated cloud service utilizing both private and public cloud to perform the distinct functions within the same organization. Therefore, an organization can maximizes their efficiencies by employing public cloud services for all non-sensitive operations, only relying on a private cloud where they require it and ensuring that all of their platforms are seamlessly integrated.

II. LITERATURE REVIEW

According to [1], Cloud computing is an emerging technology getting used in every area. Conventional organization use IT infrastructure, which isn't scalable consistent with their requirement. Associations moving their responsibility to cloud for improving their exhibition, adaptability and furthermore for lessening cost. Cloud computing is employed for deploying the hospital management system available anywhere and at any time. Here, the administrator performs the action on three modules i.e. doctor, patient and rooms allocation, where the administrator can view and access the small print. Generally, there are many public cloud computing providers like AWS, IBM Smart Cloud, GCP, and lots of others. This proposed model uses GCP because it is rising cloud computing platform with sorts of services like storage technologies, various quite databases, secure networking technologies, machine learning platforms, computing capabilities and hosting of application.

According to [2], Cloud computing is rapidly becoming a widespread alternative to costly on-premise infrastructures for delivering computing services generally and specifically for data processing services. Bearing this in mind, it's fairly convenient, to propose an architecture for the deployment of knowledge Mining services that might allow the underlying computing platform to be abstracted, leaving out of consideration of the cloud provider, technology or the supporting architecture, and that specialize in service and his flexible description, composition and deployment. For this reason, a stage for the sending of knowledge Mining services referred to as OC2DM: Open Cloud Computing data processing has been designed.

According to [3], they proposes a model of sentiment analysis of varied features of various companies' mobile phones and their overall rating. Before buying a phone, customers usually search for reviews to make a decision which phone to shop for . The model proposed during this paper provides an optimal solution for the customer for creating this decision more efficiently. during this model, each feature of a mobile is rated supported popular opinion and an overall rating for every phone is provided. Amazon is one among the most important Internet retailers, which makes way for many public reviews on their products. These reviews are collected as a sort of an open source platform and used because the dataset during this model. The gathered data is preprocessed then separated into two different sets – Training Set and Testing Set which are wont to train and test the supervised machine learning algorithms for classification. 15 commonest features of the mobile phones supported public reviews are selected from the training data set and used because the feature set during this model. Different algorithms which include Naïve Bayes, Support Vector Machine, Logistic Regression, and Stochastic Gradient Descent algorithms are utilized in this model and therefore the comparison of their performance is shown. This model provides a rating of every feature and a mean rating of the mobile supported sentiment polarity. Thus, this research work can assist potential customers to settle on the simplest product supported the opinion of the opposite users.

According to [4], Sentiment analysis is one among the fastest spreading research areas in computing , making it challenging to stay track of all the activities within the area. They present a client criticism audits on item, where we use assessment mining, text mining and slants, which has influenced the encircled world by changing their assessment on a chose item. Data utilized in this study are online product reviews collected from Amazon.com. They performed a comparative sentiment analysis of retrieved reviews.

ISSN (online):2396-9889
INTERNATIONAL RESEARCH JOURNAL OF TECHNOLOGY AND APPLIED SCIENCE
http://www.irjtas.com
(An ISO Certified Journal)
VOL. 05 Issue 02 FEBRUARY 2021

This research paper provides you with sentimental analysis of varied smart phone opinions on smart phones dividing them Positive, Negative and Neutral Behaviour.

III PROBLEM DEFINITION

Problem Using Cloud

The cloud depends on the Internet Protocol (IP), so for an application to be thought of, it should utilize IP as its correspondence instrument. While there are numerous conventions that can be run over IP, the utilization of Transport Control Protocol (TCP) is liked. Obviously the security issue has assumed the main part in frustrating Cloud figuring. Without question, putting your information, running your product at another person's hard circle utilizing another person's CPU seems overwhelming to many.. Well-known security issues such as data loss, phishing, botnet (running remotely on a collection of machines) pose serious threats to organization's data and software because in cloud every time we connect to the virtual machine a different IP address machine will allocate.

Problem in Sentiment Analysis

There are several challenges in judging sentiments form reviews, comments etc. Ordinarily in surveys there is conflicting and sporadic information. Individuals have different methods of communicating notions; sometime they use shorthand and lots of abbreviations. Usually they cannot use proper grammar in reviews. We judge positive or negative opinions from reviews using opinion words and phrases are usually used to express. These phrases and opinion words may be used in positive and negative situations. For instance great is for positive and bad is for negative. Judgment of positive and negative sentiments from review depends on context of what is around it. There are very less words that will always attach a positive or negative sentiment to an expression. Comments and reviews also contain irony and hidden emotions. The task of judging sentiments is also a challenging task, due subjective sentences and also ambiguity naturally found in opinionated text. Vagueness words are similar significance words which come in more than one time in same sentence. Ambiguity becomes a serious problem when it come with irony and convey words. So its very important to extract feature before training a machine learning model.

IV PROPOSED WORK

In these we propose a private cloud preferred over others because cloud service is provided by third party providers so for security reason private cloud give better security than others because the connection between user and virtual machine is secured by ssh. And on private cloud we can easily scale the storage and processing power at any time whenever application required. The protected, high-accessibility Web application is going. At the point when the application should be refreshed, the virtual machine pictures can be refreshed, replicated across the improvement chain, and the whole framework can be redeployed.

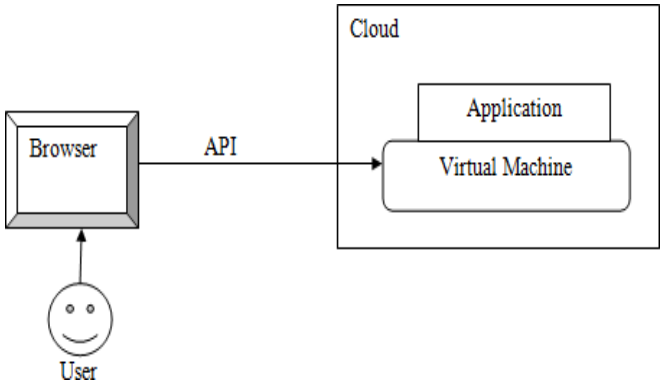

Figure 1. Proposed Block Diagram

Step-1. First user will access the browser and login into the cloud service website.
Step-2. After Successful login a secure ssh connection has been developed between user browser and virtual machine on cloud through API call.
Step-3. We can deploy an machine learning application on virtual machine over cloud.

Since machine learning algorithms work only with fixed-length vector of numbers instead of raw text, the input (in this case text data) got to be parsed. There are various methods for transforming the texts into features. In these we've taken logistic regression algorithm for amazon review analysis. it's a classification not a regression algorithm. it's wont to estimate discrete

ISSN (online):2396-9889
INTERNATIONAL RESEARCH JOURNAL OF TECHNOLOGY AND APPLIED SCIENCE
http://www.irjtas.com
(An ISO Certified Journal)
VOL. 05 Issue 02 FEBRUARY 2021

values (Binary values like 0/1, yes/no, true/false) supported given set of independent variable(s). In basic words, it predicts the likelihood of event of an event by fitting information to a logit work. Hence, it's also referred to as logit regression. Since, it predicts the likelihood; its yield esteems lies somewhere in the range of 0 and 1 (true to form)

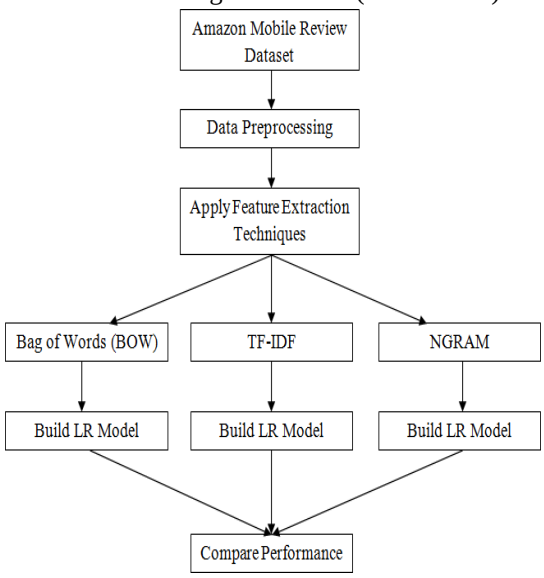

Figure 2. Proposed Block Diagram for Text Classification

Algorithm Steps

Step-1. Loading Dataset: First we will collect the amazon mobile review dataset from web and stored it for processing.

Step-2. Data Preprocessing: In these we will remove the missing fields records from the dataset and also we remove the neutral reviews, Not the dataset is merely consist positive and negative review.

Step-3. Feature Extraction: After pro-processing we apply various feature extraction techniques for training a model.

Step-4. Building a Model: we will build three Logistic Regression Model supported different feature extraction techniques.

Step-5. Compare the three model on the idea of test accuracy.

V EXPERIMENTAL ANALYSIS

For experiment we will take an virtual machine on cloud preferred over others because cloud service is provided by third party providers (Google cloud platform (GCP)) so for security reason private cloud give better security than others because the connection between user and virtual machine is secured by ssh. And on private cloud we will easily scale the storage and processing power at any time whenever application required. Figure 3 shows the virtual machine we've taken over cloud and our machine ip address is fixed means data isn't distributed in multiple machines.

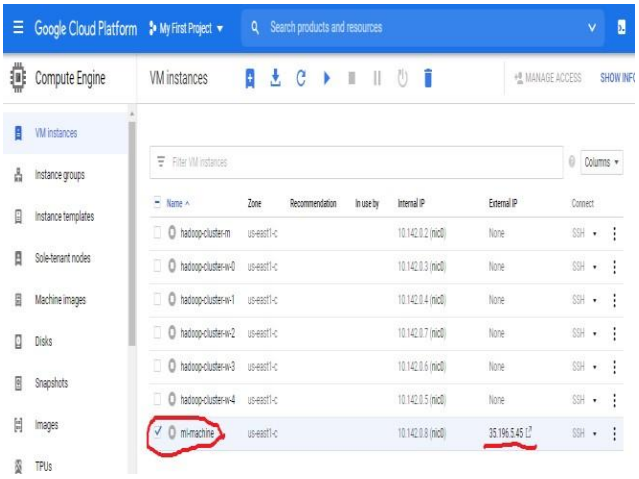

Figure 3. Virtual Machine on Cloud Environment

After clicking on start button , out virtual machine gets started and to with these we click on ssh which is nothing but creating a secure channel between browser and therefore the virtual machine terminal. ssh gives a secure login so we will securely access our resources, figure 4 shows the terminal of virtual machine on which we are starting out python environment over which we will deploying our code for classification.

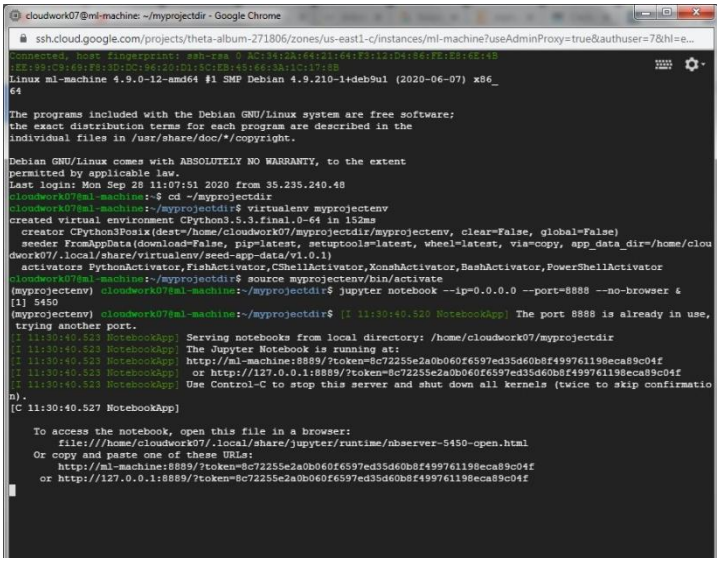

Figure 4. Starting Python over cloud computing machine

ISSN (online):2396-9889
INTERNATIONAL RESEARCH JOURNAL OF TECHNOLOGY AND APPLIED SCIENCE
http://www.irjtas.com
(An ISO Certified Journal)
VOL. 05 Issue 02 FEBRUARY 2021

After that we will connect the python and jupyter notebook (python IDE for Machine Learning) through browser employing a secure ip and port number. once we undergo the URL it can invite secure token or password using which we will access other notebooks so after giving it the secure token we will connect with the jupyter notebook, First we will load the amazon mobile review datasets which is shown in figure 5.

```
import pandas as pd
import numpy as np

df=pd.read_csv('Amazon_Unlocked_Mobile.csv')

df=df.sample(frac=0.1,random_state=10)
df.head()
```

	Product Name	Brand Name	Price	Rating	Reviews	Review Votes
394349	Sony XPERIA Z2 D6503 FACTORY UNLOCKED Internat...	NaN	244.95	5	Very good one! Better than Samsung S and iphon...	0.0
34377	Apple iPhone 5c 8GB (Pink) - Verizon Wireless	Apple	194.99	1	The phone needed a SIM card, would have been n...	1.0
248521	Motorola Droid RAZR MAXX XT912 M Verizon Smart...	Motorola	174.99	5	I was 3 months away from my upgrade and my Str...	3.0
167661	CNPGD [U.S. Office Extended Warranty] Smartwat...	CNPGD	49.99	1	an experience i want to forget	0.0
73287	Apple iPhone 7 Unlocked Phone 256 GB - US Vers...	Apple	922.00	5	GREAT PHONE WORK ACCORDING MY EXPECTATIONS.	1.0

```
df.shape
```

```
(41384, 6)
```

Figure 5. Loading Amazon Mobile Review Dataset

Data Preprocessing

After loading the info set we will explore the info and than we start processing the data, we will remove the missing records from the dataset and also the dataset contains review rating between 1 and 5. we will convert the records into positive reviews who's rating is 4 or 5, Negative review who's rating is 1 or 2, and take away the neutral reviews from the datasets who has review rating is 3, the preprocessing steps are shown in figure 6.

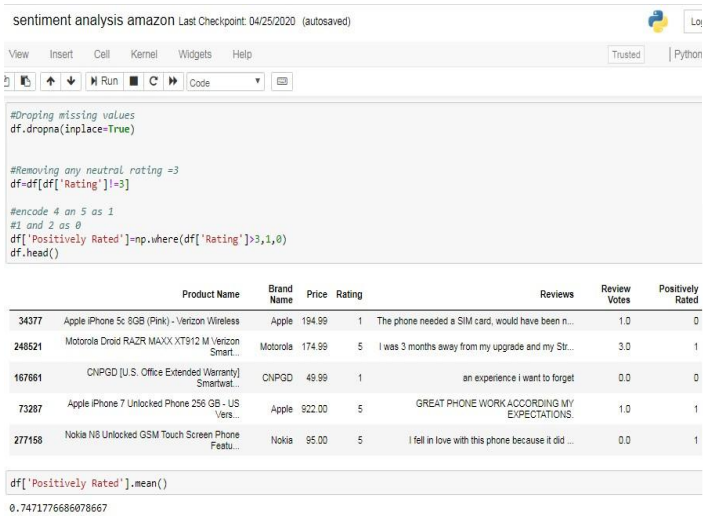

Figure 6. Preprocessing of Data

Build a Model

After preprocessing we will split the dataset into training and testing dataset and on training data we will apply various feature extraction techniques on which we will trained our predictive LR model. First Bag of Words (court vectorizer) techniques allows us to use bag of words approch by converting collection of text documents in to a matrix of token counts

First we start the tally vectorizer and fit it to our preparation information. Fitting the count vectorizer consists of the tokens of the training data and building of the vocabulary

Fitting the count vectorizer tokenizes each document by finding all sequences of characters that's numbers or letters seperated by word boundaries converts every thing to small letter and builds a vocabulary using these tokens. we will build a model on the features and therefore the prediction results are shown in figure 7.

```
from sklearn.linear_model import LogisticRegression

model=LogisticRegression()
model.fit(X_train_vectorized,y_train)
```

```
C:\Users\abhishek\Anaconda2\lib\site-packages\sklearn\linear_model\logistic.p
ed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=None, solver='warn', tol=0.0001, verbose=0,
                   warm_start=False)
```

```
from sklearn.metrics import roc_auc_score

predictions=model.predict(vect.transform(X_test))
print('AUC:',roc_auc_score(y_test,predictions))
```

```
AUC: 0.8974332776669326
```

Figure 7. Prediction result of LR model on BOW

After these we will apply another TF-IDF feature extraction techniques on training dataset. tfidf allows us to weight terms how imp they're to the document high weight are given to the terms that apper to the document but dont appear often during a corpus features with low tfidf are are available use altogether documents. Model

ISSN (online):2396-9889
INTERNATIONAL RESEARCH JOURNAL OF TECHNOLOGY AND APPLIED SCIENCE
http://www.irjtas.com
(An ISO Certified Journal)
VOL. 05 Issue 02 FEBRUARY 2021

prediction results on the feature extraction techniques are shown in figure 8.

```
from sklearn.feature_extraction.text import TfidfVectorizer

# Fit the TfidfVectorizer to the training data specifiying a minimum document frequency of 5
vect = TfidfVectorizer(min_df=5).fit(X_train)
len(vect.get_feature_names())
```

```
5442
```

```
X_train_vectorized = vect.transform(X_train)

model = LogisticRegression()
model.fit(X_train_vectorized, y_train)

predictions = model.predict(vect.transform(X_test))

print('AUC: ', roc_auc_score(y_test, predictions))
```

```
C:\Users\abhishek\Anaconda2\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWar
ed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
```

```
AUC:  0.889951006492175
```

Figure 8. Prediction result of LR model on TF-IDF

After these we can apply n-gram feature techniques on training dataset and then on these data we can trained or LR model and the test prediction results are shown in figure 9.

NGRAM

```
# Fit the CountVectorizer to the training data specifiying a minimum
# document frequency of 5 and extracting 1-grams and 2-grams
vect = CountVectorizer(min_df=5, ngram_range=(1,2)).fit(X_train)

X_train_vectorized = vect.transform(X_train)

len(vect.get_feature_names())
```

```
29072
```

```
model = LogisticRegression()
model.fit(X_train_vectorized, y_train)

predictions = model.predict(vect.transform(X_test))

print('AUC: ', roc_auc_score(y_test, predictions))
```

```
C:\Users\abhishek\Anaconda2\lib\site-packages\sklearn\linear_model\log
ed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
```

```
AUC:  0.9110661794597458
```

Figure 9. Prediction result of LR model on N-Gram

Comparison

The legitimacy of the model can be noticed utilizing blunder or exactness of the model alongside—false positive‖ and —false negatives‖.

ISSN (online):2396-9889
INTERNATIONAL RESEARCH JOURNAL OF TECHNOLOGY AND APPLIED SCIENCE
http://www.irjtas.com
(An ISO Certified Journal)
VOL. 05 Issue 02 FEBRUARY 2021

$$\text{Accuracy} = \frac{\text{Number of correct prediction}}{\text{Total number of data points}}$$

Table 1. Performance Comparison

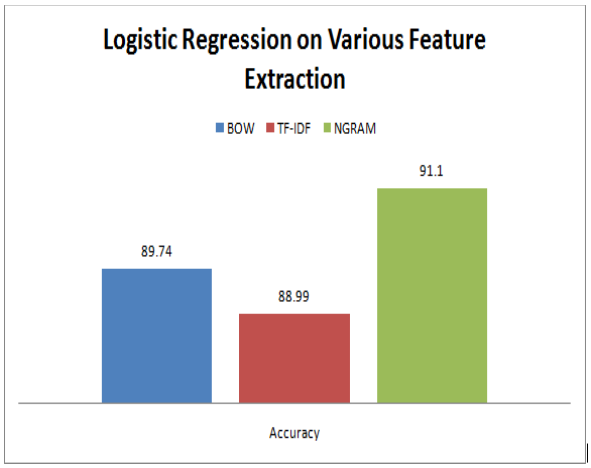

Figure 10. Comparison of Accuracy

VI CONCLUSION

The system is accurate enough for the test of reviews on amazons. For sentiment analysis we've designed our own methodology that integrates existing sentiment analysis approaches. Classification of reviews alongside sentimental analysis increased the accuracy of the system which successively provides accurate reviews to the user. In this, we will take an virtual machine on cloud preferred over others because cloud service is provided by third party providers (Google cloud platform (GCP)) so for security reason private cloud give better security than others because the connection between user and virtual machine is secured by ssh. And on private cloud we will easily scale the storage and processing power at any time whenever application required. and that we can build logistic regression model supported different feature extraction techniques like BOW (Bag of Words), TF-IDF and N-Gram , From experimental result we will say that model repose on N- Gram features provides better accuracy as compared to others.

Algorithm	Feature Extraction Technique	Accuracy
Logistic Regression	BOW	89.74%
Logistic Regression	TF-IDF	88.99%
Logistic Regression	NGRAM	91.10%

REFERENCES

[01] Ambika Gupta ; Pragati Goswami ; Nishi Chaudhary ; Rashi Bansal "Deploying an Application using Google Cloud Platform" in 2020, IEEE .

[02] Manuel Parra-Royon, Jose M. Benitez" Delivering Data Mining Services in Cloud Computing " in IEEE 2019.

[03] Abdullahil Kafi, M. Shaikh Ashikul Alam, Sayeed Bin Hossain, Siam Bin Awal, Hossain Arif" Feature-Based Mobile Phone Rating Using Sentiment Analysis and Machine Learning Approaches" in 2019, IEEE.

[4] Pankaj, Prashant Pandey, Muskan, Nitasha Soni " Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews " in IEEE 2019.

[5] Jianfeng Yang, Zhibin Chen. –Cloud Computing Research and Security Issues‖, 978-1-4244-5392- 4/10©2010 IEEE.

[6] Shuai Zhang, Shufen Zhang, Xuebin Chen, Xiuzhen Huo. –The Comparison between Cloud Computing and Grid Computing‖, 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), 978-1- 4244-7237-6/ 2010 ©IEEE.

[7] Donald Robinson ,—Amazon web services made simple‖

[8] Thomas B Winans and john seely brown,‖cloud

ISSN (online):2396-9889
INTERNATIONAL RESEARCH JOURNAL OF TECHNOLOGY AND APPLIED SCIENCE
http://www.irjtas.com
(An ISO Certified Journal)
VOL. 05 Issue 02 FEBRUARY 2021

computing –a collection of working papers‖,May2009

[9] Introduction to the cloud computing architecture white paper 1st edition 2009 by sun Microsystems

[10] Greg Boss,Padma Malladi,Dennis Quan,Linda Legregni,Harold Hall- –cloud computing‖ 8 October 2007 IBM