

A Machine Learning Model for Sentiment Analysis using Online Product Reviews

¹Om Prakash Samantray, G.Anuradha, K.SaiYeswanth,
Ch.SaiSandeep, Ch.VenkataPavan

¹ Assistant Professor, CSE, Raghu Institute of Technology, Visakhapatnam
^{2,3,4,5} Student, Raghu Institute of Technology, Visakhapatnam.

Submitted: 15-07-2021

Revised: 29-07-2021

Accepted: 31-07-2021

ABSTRACT: Sentiment analysis is defined as the process of mining of data, view, review or sentence to predict the emotion of the sentence through natural language processing (NLP). The sentiment analysis involves classification of text into different phases like “Positive”, “Negative” or “Neutral”.

It analyses the data and labels the ‘better’ and ‘worse’ sentiment as positive and negative respectively. Thus, in the past years, the World Wide Web (WWW) has become a huge source of raw data generated custom or user. Using social media, e-commerce website, movies reviews such as Facebook, twitter, Amazon, Flipkart etc. user share their views, feelings in a convenient way. In the web, millions of people express their views in their daily interaction, either in the social media or in e-commerce which can be their sentiments and opinions about particular thing.

These growing raw data are an extremely high source of information for any kind of decision making process either positive or negative. To analyse such huge data automatically, the field of sentiment analysis has turned up. Therefore, to find polarity or sentiment of customers there is a demand for automated data analysis techniques. In this paper, a sentimental analysis of product review data is performed using machine learning algorithms. The algorithms used in this experiment include, Support Vector Machine (SVM) and Naive Bayes classifiers. A comparative study of their performances is also presented.

KEYWORDS: Sentiment Analysis, Product Review, Machine Learning,

I. INTRODUCTION

Sentiment is a feeling-driven attitude, idea, or judgement. Sentiment analysis, often known as opinion mining, investigates people's feelings about various entities. The internet is a very useful tool when it comes to sentiment data, this is where you should go. People can upload

their own content on various social media platforms, such as forums, microblogs, and online social networking sites, from the perspective of the user. From the standpoint of a researcher, many social media sites expose their application programming interfaces (APIs), allowing researchers and developers to collect and analyse data.

For example, Twitter now offers three APIs: the REST API, the Search API, and the Streaming API. Developers can acquire status data and user information via the REST API; the Search API allows them to query specific Twitter content, while the Streaming API allows them to collect Twitter content in real time. Developers can also mix and match APIs to construct their own apps. As a result, sentiment analysis appears to have a solid foundation based on huge web data.

However, there are various limitations in this form of internet data that could make sentiment analysis difficult. The first fault is that, because people are free to write whatever they choose, the quality of their opinions cannot be guaranteed. Online spammers, for example, instead of providing topic-related opinions, post spam on forums. Some spams are completely pointless, while others contain irrelevant or misleading opinions. The second problem is that such internet data does not necessarily have a ground truth. A ground truth is more akin to a label for a certain viewpoint, stating whether it is positive, negative, or neutral.

One of the datasets with ground truth is the Stanford Sentiment 140 Tweet Corpus, which is also open to the public. There are 1.6 million machine-tagged Twitter messages in the corpus. Each communication is assigned a tag depending on the emoticons found inside it (as ☺ positive or ☹negative).

The data used in this paper is a collection of product reviews gathered from Amazon over a

period of time. The issues listed above have been partially addressed in the following two ways: Before a product review can be published, it must first pass the inspection.

Second, each review must provide a rating that may be used as the basis for comparison. The ranking is based on a five-star system, with five stars being the highest. This work addresses a key issue in sentiment analysis: sentiment polarity categorization.

Motivation of this work is the importance of opinions of people on a product. Other people's opinions about a product can have an impact on our decision. People used to gather information about a product through direct sources such as friends, relatives, consumer reports, and strangers in the past. We now have various options for looking at other people's viewpoints (ex: Internet). The internet makes it possible to gather opinions from people all around the world. People use e-commerce sites such as Amazon, eBay, Flipkart, Snapdeal, and others to find product reviews and comments, and then buy the product based on the reviews. Because social media has such a large influence, it is now also used to learn about products.

Some companies employ surveys, opinion polls, and social media to get feedback on their products. Because social media has such a large influence, it is now also used to learn about products. Some companies employ surveys, opinion polls, and social media to get feedback on their products. Because there may be thousands of reviews and varied opinions on a given product, it is impossible to look through all of them in order to make a judgement, which is where sentiment analysis comes in. Sentiment analysis considers and analyses the words included in the evaluations to categorise them into different levels such as good, poor, worst, and average.

Objectives of this work are, reviewing data, sentiment phrase extraction, part of speech tagging, sentiment phrase identification, sentiment score computation, feature vector construction and sentiment polarity categorization.

II. LITERATURE SURVEY

The categorization of sentiment polarity is a fundamental problem in sentiment analysis. The task is to categorise a piece of written language into one of two sentiment polarities: positive or negative (or neutral). There are three stages of emotion polarity categorization, depending on the breadth of the text: document level, sentence level, and entity and aspect level. The document level is concerned with whether a document communicates

negative or positive sentiment as a whole, whereas the sentence level is concerned with the sentiment categorization of each sentence; the entity and aspect level then focuses on what people like or dislike from their perspectives. We will just evaluate some past work, on which our research is largely based, in this part because reviews of much work on sentiment analysis have already been included. Based on customer reviews, Hu and Liu [1] summarised a list of positive terms and a list of negative words, respectively. There are 2006 words on the good list and 4783 terms on the negative list. Both lists also include some commonly misspelt words found in social media postings.

Sentiment categorization is a classification problem in which features containing views or sentiment information must be found prior to classification. Pang and Lee [2] proposed that for feature selection, objective sentences be removed and subjective sentences be extracted. They presented a text-categorization technique that uses the minimum cut to identify subjective content. Gann et al. chose 6,799 tokens based on Twitter data, assigning each one a sentiment score, the TSI (Total Sentiment Index), indicating whether it is a positive or negative token. Y Xue et al. [3] proposed a pair of generative models, MaxEnt-JABST and JABST, that extracted fine-grained opinions and aspects from reviews (online).

The sentiment polarity, as well as specific and general thoughts and attributes, were retrieved using the JABST model (SP). Additionally, the MaxEnt-JABST architecture included a maximal entropy classifier for more precisely differentiating aspects or opinion words. These designs were evaluated numerically and qualitatively in terms of restaurants and electrical equipment.

Manvee Chauhan et al. [4] presents an overview of product reviews by categorising them as good, negative, or neutral.

The amount of information available on the internet is enormous. Because reviews are highly unstructured, machine learning methodologies such as naive Bayes and support vector machine algorithms are used. These algorithms start with unstructured product reviews, perform pre-processing, calculate polarity of reviews, and extract features for experiment. They used ML algorithms for experiment. Mohammad Shabaz et al. [5] devised a new formula that runs from -1 to +1 to obtain the sentiments count, where negative values signify negative sentiments, positive values denote positive sentiments, and 0 denotes neutral sentiments. We may assume that the results we acquired are close to accurate with this method because there is no standard for

accuracy and it varies from person to person, therefore conducting performing sentimental analysis always gives approximate results.

Prabha PM Suryaet. al. [6] used the Naive Bayes classifier for classification and semantic analysis. The dataset used in their study is the Amazon product review dataset from the University of California at Irvine's repository. The dataset contains approximately 600 records, each of which is examined using the Nave Bayes approach, which is a probabilistic approach that yields a matrix as a result.

Prashast Kumar Singhet. al. [7] automated the process of collecting online, end user reviews for any given product or service and analysing those reviews for attitudes expressed regarding specific characteristics. This entails removing irrelevant and unhelpful reviews, as well as quantifying the sentiments of thousands of (good) reviews. Looking at the above related works, we have decided to use product review data for sentimental analysis using machine learning algorithms. Machine learning algorithms are very useful for

sentimental analysis, malware analysis and other classification problems [8]. Best feature selection algorithms can be used to increase the efficiency of the proposed model [9]. The sentimental analysis can also be used with TF/IDF method because the product reviews involves textual data and the said method works best with textual data [10].

III. SYSTEM OUTLINE

The data was acquired from Amazon and included reviews for laptops, cameras, mobile phones, tablets, video surveillance, and televisions. Following that, stemming, stop word removal, and punctuation mark removal were carried out, and the text was transformed into a bag of words.

This dataset was compared to opinion lexicons, which included 4783 negative and 2006 positive words, as well as sentiment scores for each sentence. The NB and SVM were used in conjunction with score and other features, and various accuracy was computed. The system outline is shown in figure 1.

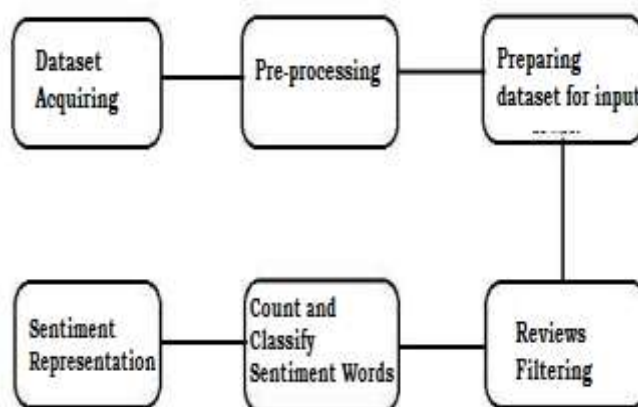


Figure 1: System outline of sentiment analysis of product review data.

Data is collected from online sources like Amazon where user provide their opinions on products. Importing positive and negative datasets from a file and integrating them into a single file is called data filtering. There could be a lot of unnecessary symbols and numbers in the data sets. To improve efficiency, these factors must be fixed or solved. As a result, the undesirable symbols and numbers are deleted during this process.

Extraction of all comparable terms (feature words) such as adjective, adverb, and verb from the datasets in the file. The datasets are then labeled as "pos" for positive and "neg" for negative, respectively. Then, based on the frequency distribution of the gathered words, 5000 words were chosen for training. For improved training,

the data is once again shuffled using a random seed. The labeled datasets are separated into training and testing percentiles of 70 and 30 percent, respectively. The user can test and analyse the respective model in the testing model by preparing the input data. The symbol and number are removed as part of the preprocessing. Using stored features to map to user input (based on training dataset). The data is then fed into the saved model for prediction.

IV. EXPERIMENTAL RESULTS

The experiment is performed using Python programming language because it has rich set of build in packages and features to carry out machine learning classifications efficiently. NumPy

(Numerical Python) is a Python library with high-level functions for manipulating arrays mathematically. It is made up of multidimensional array objects and a collection of array processing techniques. The following operations can be carried out with NumPy: Operations that are both mathematical and logical. Fourier transformations and routines are used to manipulate shapes.

The Bayes theorem is used to create the Naive Bayes classifier. It makes a significant assumption about independence. It's sometimes referred to as a "independent feature model." It

assumes that the existence or absence of a certain class characteristic has no bearing on the presence or absence of any other class feature. A supervised learning model can also be used to train the Naive Bayes classifier. It uses the highest similarity method. The variance of the variable must be calculated for each class separately, rather than for the entire matrix. When the number of inputs is large, Naive Bayes is utilized. It produces output that is more sophisticated. The Prediction table shows the probability of each property. The Bayes theorem is shown in figure 2.

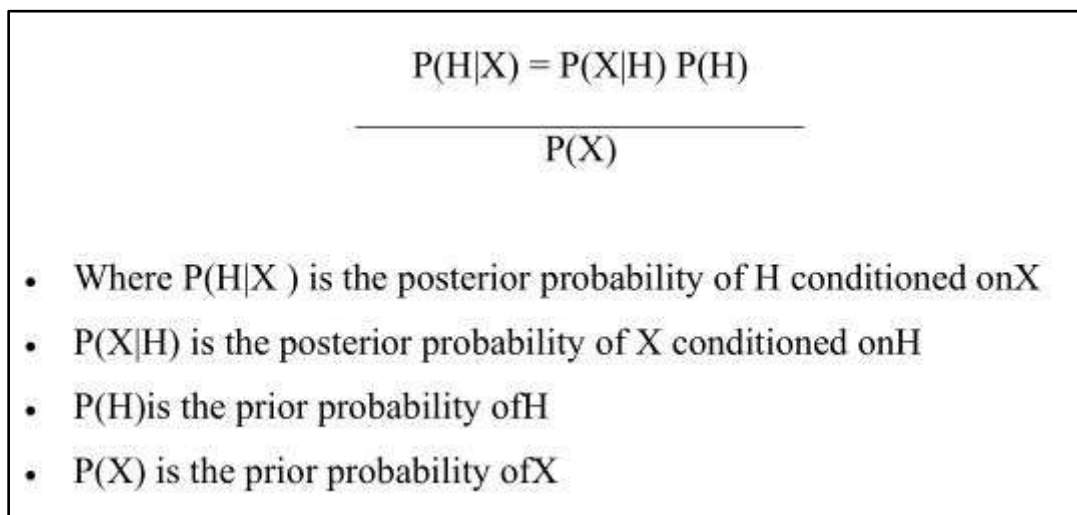
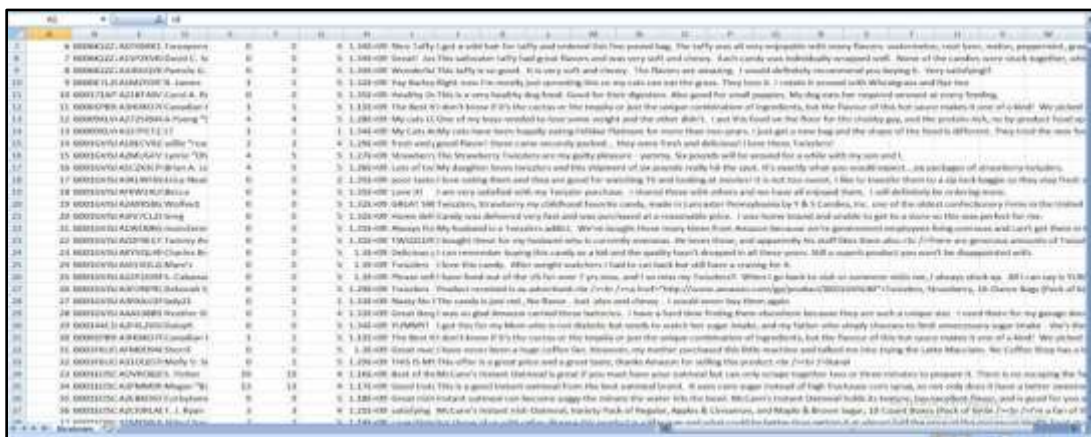


Figure 2: Bayes Theorem

SVM (Support Vector Machine) is a supervised machine learning technique that can be used to solve classification and regression problems. It is, however, mostly employed to solve categorization difficulties. Each data item is plotted as a point in n-dimensional space (where n is the number of features you have), with the value of

each feature being the value of a certain coordinate in the SVM algorithm. These two algorithms found suitable for many machine learning related real-world classification problems [9]. Therefore, we have decided to use them in this work. The dataset used in this experiment is shown in figure 3.



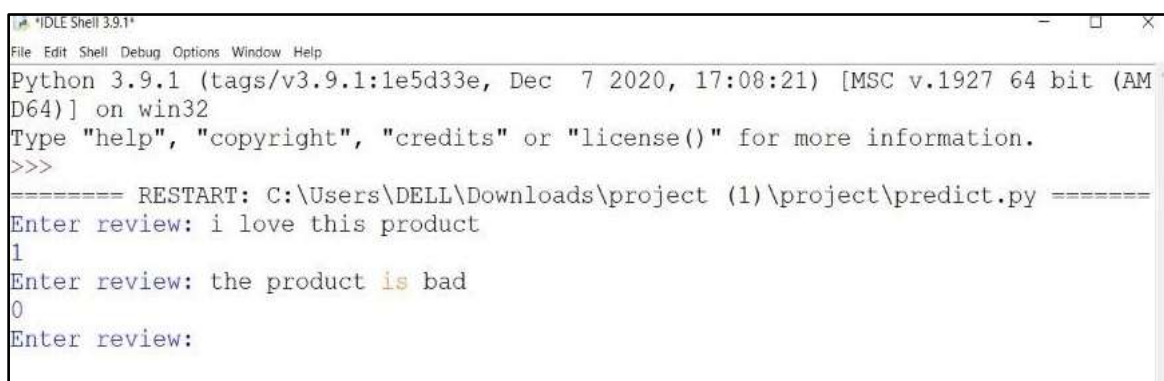
ID	Sentiment	Text
1	Positive	My laptop got a water leak but the laptop and keyboard are fine. The laptop is all very responsive with many features, excellent, fast, easy, medium, professional, great...
2	Positive	For the software, it's very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
3	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
4	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
5	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
6	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
7	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
8	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
9	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
10	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
11	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
12	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
13	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
14	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
15	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
16	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
17	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
18	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
19	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
20	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
21	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
22	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
23	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
24	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
25	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
26	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
27	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
28	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
29	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
30	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
31	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
32	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
33	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
34	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
35	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
36	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
37	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
38	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
39	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...
40	Positive	My laptop is very easy to use and very fast and very easy to use. Each card is very individually oriented well. None of the cards were stuck together, and...

Figure 3: Dataset for sentiment analysis

We only receive adjective terms via filtering by deleting stop words from the reviews, because there are many reviews for a certain product, we only require the adjective words in those reviews, and the remaining nouns and pronouns need to be ignored, thus we utilise Stop words to ignore such terms, noun and pronoun words are included in the stop words variable. We only receive adjective terms via filtering by deleting stop words from the reviews. Our first goal is to count the number of words in our output, which is a bar graph that displays customer sentiments based on the words used in their evaluations. To assist with the count, we used numerous software. In the previous step we removed

all stop words and reviews are filtered so that can be used in this step. Once we have all of the essential adjectives from the previous stage, we break them into positive and negative categories, and within those positive and negative categories, we further separate them into good, best, anger, bad, and other categories based on the customer's expressed emotions. We even figured out how many sentiment words were in each line.

After applying the ML algorithms we have classified the textual reviews into binary classification indicated by zero (Negative review) and one (positive review). A sample output of our experiment is depicted in figure 4.



```

Python 3.9.1 (tags/v3.9.1:1e5d33e, Dec 7 2020, 17:08:21) [MSC v.1927 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: C:\Users\DELL\Downloads\project (1)\project\predict.py =====
Enter review: i love this product
1
Enter review: the product is bad
0
Enter review:
  
```

Figure 4: Sample output of Semantic analysis using Machine Learning

The performance metrics used in our experiment are accuracy, False Positive Rate (FPR), recall and

precision. The values achieved from the experiment for the above metrics are given in table 1.

Table 1: Results of the experiment

Algorithm/ Metrics	Accuracy	FPR	Recall	Precision
NB	0.93	0.038	0.966	0.976
SVM	0.98	0.003	0.974	0.997

From the experiment it is observed that SVM has better accuracy in the sentimental analysis on the selected dataset. The accuracy obtained by SVM is 98% which is better than the accuracy of NB algorithm. The False positive rate of the SVM algorithm is also very less which proves this algorithm as the best method for semantic analysis in our experiment.

V. CONCLUSION

With the number of items available increasing all the time, deciding on one is becoming increasingly challenging. As a result, the demand for sentimental analysis is rapidly expanding. Despite the fact that sentimental analysis tasks are difficult due to their natural language processing origins, substantial progress has been made in recent years due to increasing

demand. Not only do customers want to know about a product, but firms also want to know how their product is doing on the market. Sentiment analysis and opinion mining will remain relevant for the foreseeable future due to the increased demand for product insights and the technical obstacles that the sector is now confronting. Opinion mining systems of the future will require a stronger link between extensive information bases and reasoning processes influenced by human mind and psychology. This will result in a better understanding of natural language opinions and a more efficient bridge between unstructured data in the form of human thinking and structured data that can be evaluated and processed by a machine. Intelligent opinion mining algorithms can handle semantic knowledge, make analogies, learn continuously, and detect

emotions, all of which contribute to very efficient sentiment analysis. The ML techniques used in our experiment produced positive results in categorising product reviews. For almost all product related reviews, SVN received 98% accuracy and NB received 93% accuracy.

REFERENCES

- [1] Hu, Mingqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004.
- [2] Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." arXiv preprint cs/0409058 (2004).
- [3] Xue, Yingbin, Xiaoye Wang, and ZanGao. "Multi-classification Sentiment Analysis Based on the Fused Model." 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2019.
- [4] Chauhan, Manvee, and DivakarYadav. "Sentimental analysis of product based reviews using machine learning approaches." Journal of Network Communications and Emerging Technologies (JNCET) 5.2 (2015): 19-25.
- [5] Shabaz, Mohammad, and Ashok Kumar. "AS: a novel sentimental analysis approach." International Journal of Engineering and Technology 7.2.27 (2018): 46-49.
- [6] Surya, Prabha PM, and B. Subbulakshmi. "Sentimental Analysis using Naive Bayes Classifier." 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN). IEEE, 2019.
- [7] Singh, Prashast Kumar, et al. "An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites." 2014 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence). IEEE, 2014.
- [8] Samantray, Om Prakash, Satya Narayan Tripathy, and Susanta Kumar Das. "A Data Mining Based Malware Detection Model using Distinct API Call Sequences." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8.7 (2019).
- [9] Samantray, Om Prakash, and Satya Narayan Tripathy. "A Knowledge-Domain Analyser for Malware Classification." 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA). IEEE, 2020.
- [10] Samantray, Om Prakash, and Satya Narayan Tripathy. "IoT-Malware Classification Model Using Byte Sequences and Supervised Learning Techniques." Next Generation of Internet of Things. Springer, Singapore, 2021. 51-60.