Passer Journal

Volume 2 | Issue 2 Article 6

17-11-2020

# Supervised Sentiment Analysis Model of Textual Content for Images

Wrya Anwar Hayder
*Department of IT , College of Computer & IT, University of Garmian, Kalar, Kurdistan Region, Iraq*



## Recommended Citation

Hayder, Wrya Anwar (2020) "Supervised Sentiment Analysis Model of Textual Content for Images," *Passer Journal*: Vol. 2 : Iss. 2 , Article 6.
DOI: 10.24271/psr.16
Available at: https://passer.garmian.edu.krd/journal/vol2/iss2/6

Available online at www.garmian.edu.krd

Passer Journal

Passer 2 (2020) 81- 86

http://passer.garmian.edu.krd/

# Supervised Sentiment Analysis Model of Textual Content for Images

Wrya Anwar Hayder *

*Department of IT , College of Computer & IT, University of Garmian, Kalar, Kurdistan Region, Iraq*

Received 14 August 2020; revised 23 September 2020;
accepted 28 October 2020; available online 17 November 2020

doi:10.24271/psr.16

## ABSTRACT

Sentiment analysis is a domain in machine learning that tries to analyze people's emotion, feeling, opinion and attitudes towards particular service or product. It aims to extract feelings and opinion from textual reviews; therefore, it is closely related to natural language processing (NLP). Social media has provided a huge amount of text reviews, which is practically impossible to read and analyze the emotions, attitudes and opinions that were expressed in those textual data. Sentiment analysis is a machine learning concept to classify a textual data according to reviewers' emotion and attitudes about a service or product, which helps in determine strong or weak production. In this paper work we aim to develop a sentiment analysis model of texts for images. Different machine learning algorithms are tested such as Naive Bays, Logistic Regression and Support Vector Machine (SVM), in order to develop a high accuracy sentiment analysis system. The model is developed to determine whether a text has positive or negative emotion for images. The outcome of the project work shows that SVM algorithm has a better performance for such purpose, while Logistic Regression algorithm shows a faster execution time.



*Keywords: Machine learning, sentiment analysis, NLP model, Sentiment system, Machine learning model, Text mining.*

## 1. Introduction

Machine learning is a branch of computer AI, it is a science of using machine to acquire new knowledge from data and information. Machine learning recently is used in data science, natural language processing, biometric, medical diagnostics, and detection of credit card fraud, games and robotics.

Simply machine learning can be represented as computer program model that trained to best maps input variable (x) to an output variable (y) as y=f(x). The task of the learning function model is to predict the output (y) of any given input (x) [1].

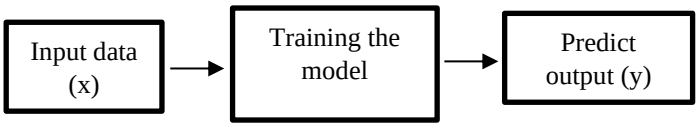


**Figure 1:** Machine Learning Model

Nowadays, people share their emotion, feeling and opinion in the form of cartoons, memes and images as well as facial expressions [2]. The process of classifying these emotions and feelings is known as sentiment analysis [3]. Sentiment analysis is a machine learning technique to extract and predict emotions or opinions of the people from a dataset [4].

Sentiment analysis has seen a significant importance in recent years by many companies to classify whether customers or users are satisfied or not about service or product. It is also useful in understanding a general trend in people emotion, opinion and attitude through their comments and posts. This process has become more influential with the increased usage of social media [5].

Even though many researchers have been applying sentiment analysis for the purpose of movie reviewing and social media commenting, applying sentiment analysis models for text that embedded with images has not been covered [6]. Since these texts are shorter and have a different grammatically structure, it is important to testify machine learning sentiment analysis models for such purpose. As part of this paper work, we aim to classify and predict emotion or feeling of images that have text embedded within them. The classification category falls into two form which they are positive or negative feeling or emotion. We will testify three most common machine learning classification algorithms in order to propose a suitable prediction system.

The main concepts, models and algorithms to develop a sentiment analysis system are mentioned in the following section. It is divided into three main sub sections; first we review the main concept of machine learning concept. Then we cover the existing

* Corresponding author
*E-mail address:* wrya.anwar@garmian.edu.krd (Instructor).
Peer-reviewed under the responsibility of the University of Garmian.

major learning frameworks and algorithms, finally we mention the natural language process (NLP) and sentiment analysis application.

### 1. 1 Machine Learning

Machine learning is a field of artificial intelligence that aims to develop computer algorithms model that learn from large complex data sets to produce meaningful output. Different machine learning models and techniques may be required for different applications and domains. It depends on whether the application is for interpreting complex data or simply used for classification and prediction [1, 6, 7]. In the next section we try to review the most common machine learning models and algorithms.

### 1. 2 Learning Frameworks

There are three main framework of learning that can be applied by a machine, these frameworks are:

#### 1. 2. 1 Supervised Learning:

In supervised learning the input data (x) has well-known labels, such as positive text and negative text. The model is prepared by a training process based on predefined features. For example positive and negative words are the main text features for the model. The model will be trained by calculating the total of positive or negative word in sentence. If the total of positive words is greater than negative then the model will predict the output (y) to be positive sentence. If the total of negative words for a sentence is greater than positive the model prediction of (y) will be negative. The main machine learning model used for supervised learning is:

- Classification and Regression

Classification is the process of mapping input (x) to a correct class of output(y) for (categorical) prediction. It is widely used in every day process, such as determined whether an email is spam or non-spam. Reviews of movies can be classified regarding the reviewer's opinion as positive or negative. In addition, it could be used in medical diagnostics by classifying an image as correct type of diseases.

Regression on the other hand, is usually used for (continuous) prediction training models, for instance predicting weather based on previous data can be referred to as regression process.

In this paper we emphasize in text classification, which is used in many applications nowadays. Text classification is the process of assigning a text known as document (D) to one of pre-defined classes known as (C) where C={c1,c2,…ci}. Many algorithms models of machine learning can be applied for the purpose of classification and regression, the most common algorithms are:

### i.Naïve Bayes

It is the most common and most widely used by researchers because of its simplicity in training and classification process. It is based on probability classification which can learn from as set of document by calculating the maximum probability of word presence within a sentence. It works by given document d the class c* = arg maxc P(c | d). The main classifier rule is:

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)}$$

Despite its simplicity naïve base-based text classification still perform surprisingly well.

### ii.Support Vector Machine (SVM)

Another algorithm that has shown to be highly effective at text classification process is support vector machine. It depends on the large-margin rather than probability. Different from Naïve Base SVM basic technique is to find a hyperplane vector that used to separates the document in a binary classification. Then the task is to determine the margin between two classes, in which the margin is far from any document.

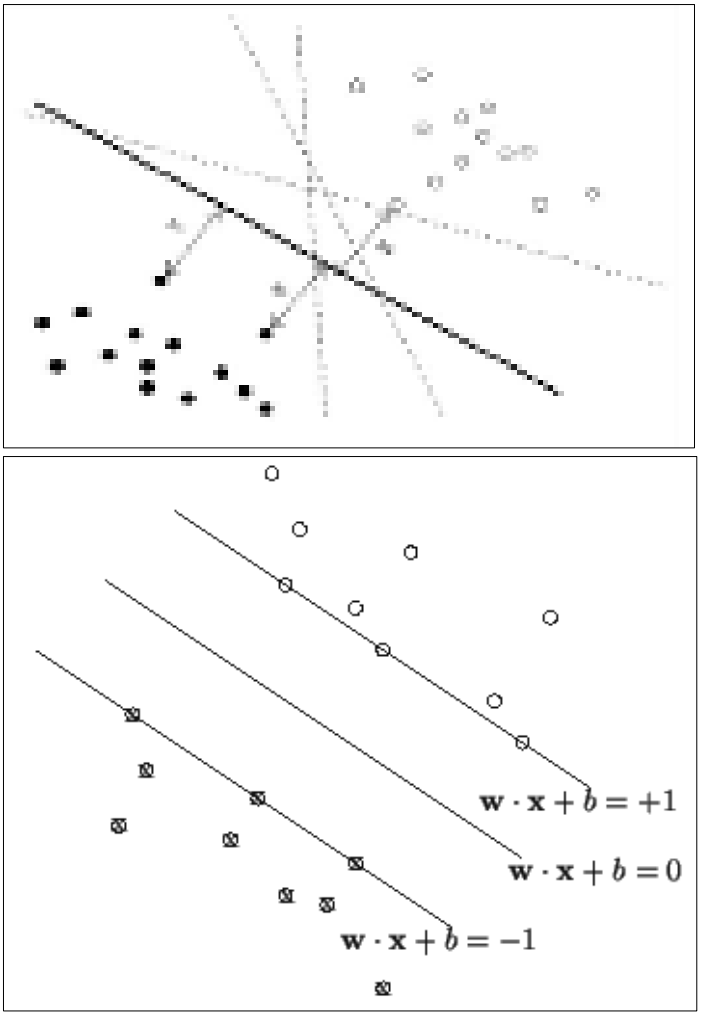


**Figure 2:** SVM Hyperplane Vector

### iii.Logistic Regression

Another most famous classification and regression algorithm is logistic regression. The main concept work is very much like linear regression, but the logistic function is a sigmoid function which takes values from 0 to 1.

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The above algorithms of supervised learning the most common and widely used algorithm used among researchers, they are fast, robust and easy to apply algorithms [6, 7, 8 , 9].

### 1. 2. 2 Other Learning Frameworks

Because the applied learning models are supervised algorithms, the introduction section covers the supervised model. In addition to the supervised, there are two other types of learning frameworks which they are (**Semi-supervised**) and (**unsupervised**) learning [10, 11]. The overview of the learning frameworks and their own algorithms are shown in the table below:

**Table 1:** Machine learning models and algorithms

| Algorithms | Supervised | Semi-supervised | Unsupervised |
|---|---|---|---|
| **Naïve Base** | Yes | Yes | No |
| **Logistic Regression** | Yes | Yes | No |
| **Support Vector Machine** | Yes | Yes | No |
| **K-Mean Clustering** | No | Yes | Yes |
| **Deep Learning- based on Neural Network** | Yes | Yes | Yes |

### 1. 3 Related Work

Study of sentiment analysis has been applied in a variety of areas such as analyzing movie reviews, social media comments and analyzing a service or product in order to improve them according to the user's need.

It has seen an increase interest among researchers for the purpose of movie review analysis for recommendation systems. It has been widely used for this purpose of move industry to understand viewer's positive reactions and proposing related movies [12]. Different supervised models applied to understand the emotion sentiment of the audience toward specific moves such as: Naïve base, SVM and logistic regression and these models showed a good prediction results which commonly upper 80% percent of accuracy [13].

Sentiment analysis applied for the purpose of social media comments, in order to understand the main stream response to specific news, public events and a specific services or product (14), as in 2012 sentiment analysis is used in USA wide public research study to measure the positive or negative affect of popular people on their followers based on their tweeter comments and how they affect the main attitude of the community [15]. Not restricted to this, it has been used by the primary Care and Public Health and center for Health policy imperial college London of United Kingdom to understand patient positive or negative descriptions of their health care in relation to the hospital clearness and services dignity and respect. The Naïve base algorithm showed better accuracy than Support Vector Machine by ration 88.6% and 84.6 respectively [16].

Carrying out sentiment analysis on product reviews has been studied by researchers to testify the algorithms for the industry purpose [17]. The analysis applied to the text comments for variety of products belong to beauty, books and electronics. The text comments of the products are analyzed using supervised machine learning algorithms to understand users and buyer positive or negative attitude towards new products, support vector machine showed better accuracy in comparison with Naïve based algorithm by 80% accuracy [18, 19].

Since the current research studies used full text with complete language syntax and grammar rules, in this study we apply sentiment analysis to extracted text quotations for images in order to understand and propose the supervised model in this specific purpose. The extracted textual content is usually shorter and have different syntax organization with usually embedded meaning.

### 1. 4 Text Processing and Sentiment Analysis Applications

Sentimental analysis is rapidly growing research filed in the area of online text comments and reviews which has been posting on different web sites where people share their opinion [20]. Text processing has been significantly used in our everyday applications. Text can be also extracted from images to be processed as natural language processing [10]. The text information constructs a huge amount of digital data over World Wide Web. Many huge business companies investing in using textual processing machine learning in their system to understand users' data and producing solutions, for instance:

Amazon: the online shopping system applies classification algorithms to produce product recommendations according to their customers' needs. Netflix a huge online move rental system is also using machine learning classification to predict and rate movies according to the customers' desires.

Sentiment analysis in fact is the process of textual classification to predict the attitude and opinion of customers and users [10, 19, 21].

## 2. Method of Proposed Model and Prototype

In this paper work we used python, image library imglib and NLTK (Natural Language Tool Kit) to extract text from images then classifying the text using Naïve Bayes, Logistic Regression and Support Vector Machine (SVM). The overall flowchart of the method is:

### i.The dataset file

The dataset used for this research is textual quotes for images available for research purpose [22]. This dataset is considered the most suitable and valid dataset currently available which related to this specific work. The dataset quotes classified into positive and negative and neutral which only a subset of strongly positive and negative sets has been chosen; in this context it refers to as document D for example (the most powerful thing in life is love), in addition each document is labeled with a fixed class referred to as C in this context the classes are (positive or negative). So the training set of D and C will look like: (D1,C1),(D2,C2),…(Di,Cj).

### ii.Data cleaning:

Removing punctuation and stop words for instance I, Me, My, and …. etc.

### iii.Fetch document words and create frequency distribution:

After data cleaning the most frequently occurring words will be stored in a table called feature table [23]. This contains the words and their probability of being positive or negative for instance our algorithm top 15 feature words are:

| **Feature Words** | **Probability of being positive or negative** |
|---|---|
| contains(wasted) = True | neg : pos = 12.6 : 1.0 |
| contains(spectacular) = True | pos : neg = 7.5 : 1.0 |
| contains(stupid) = True | neg : pos = 6.5 : 1.0 |
| contains(refreshing) = True | pos : neg = 6.2 : 1.0 |
| contains(bland) = True | neg : pos = 5.9 : 1.0 |
| contains(touching) = True | pos : neg = 5.8 : 1.0 |
| contains(awful) = True | neg : pos = 5.6 : 1.0 |
| contains(providing) = True | pos : neg = 5.5 : 1.0 |
| contains(poorly) = True | neg : pos = 5.4 : 1.0 |
| contains(awfully) = True | neg : pos = 5.1 : 1.0 |
| contains(theatrical) = True | pos : neg = 4.9 : 1.0 |
| contains(progresses) = True | pos : neg = 4.9 : 1.0 |
| contains(blowing) = True | pos : neg = 4.9 : 1.0 |
| contains(mess) = True | neg : pos = 4.8 : 1.0 |
| contains(arthur) = True | neg : pos = 4.4 : 1.0 |

Training the algorithm model is done using Naive base, Logistic Regression and Support vector machine.

### iv.Applying the model to extracted text from images:

Unlike human machine cannot understand text content of an image unless it has been extracted to a text string. At the beginning step the program opens dialog box for opening image. Then by using python Optical Character Recognition (OCR) pytesseract library which detects text content on images and convert it to text that can understood by computer. Then the generated text can be processed through Natural Language processing (NLP), the algorithm is:

*Install PTL Image library*

*Import Image*

*Import pytesseract*

*Read image*

*Scan image and convert image to gray scale bitmap image*

*Convert image to matrix of black and white dots*

*Finally extract black dots which represents the text edge*

v.Match the feature table with extracted text:

Read the text of the image and extract the words that matches the feature words that has been trained by the algorithms, the model will calculate the probability of each feature word to determine whether it is a more likely to be a positive or negative text.

Finally the model can check the accuracy of the classifier performance by computing the accuracy after prediction test.

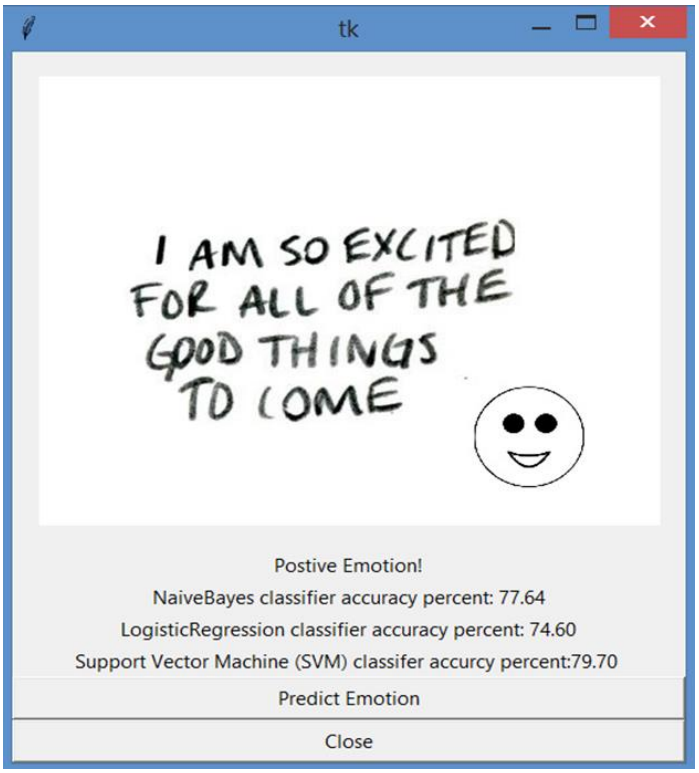


**Figure 3:** Proposed Prototype of the Sentiment Analysis using Python

## 3. Result and Discussion

In this paper work we proposed a sentiment analysis system for text that embedded within images. Since this process is a supervised classification model, we applied three most common classification algorithms which they are: **Naïve base**, **Logistic Regression** and **Support Vector Machine (SVM)**.

These algorithms have been applied previously for the purpose of movie review and social media comments. In this paper work we applied the algorithms upon images with textual content, because the text within image has a different form than movie reviews and social media comments. After testing the accuracy of these algorithms we can see that SVM showed a better performance with 79.7 that Naïve base with 77.64 and Logistic Regression with 74.6. Overall, all the three algorithms successfully predicted the text within the images.

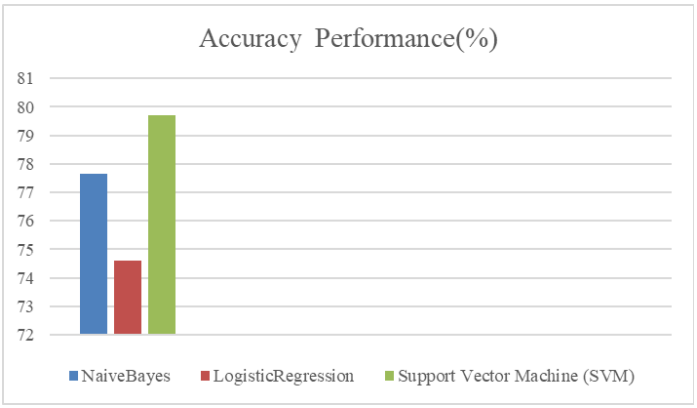


**Figure 4:** Algorithms Accuracy Performance Comparison

But as compare to the other research work about movie reviews and social media all three algorithms showed lower accuracy, that because textual comments with images are shorter and less grammatically structured.

Regarding the execution time required for each algorithm by using a laptop windows8 core i5 with 2GH CPU and 4 GB of RAM. The time has been calculated by importing python time library. As it has been shown at the figure above Logistic Regression has shown a better execution performance by 39.3 seconds, followed by Naïve Bays by 44.25 seconds, while SVM needed more execution time by 45.09 seconds.

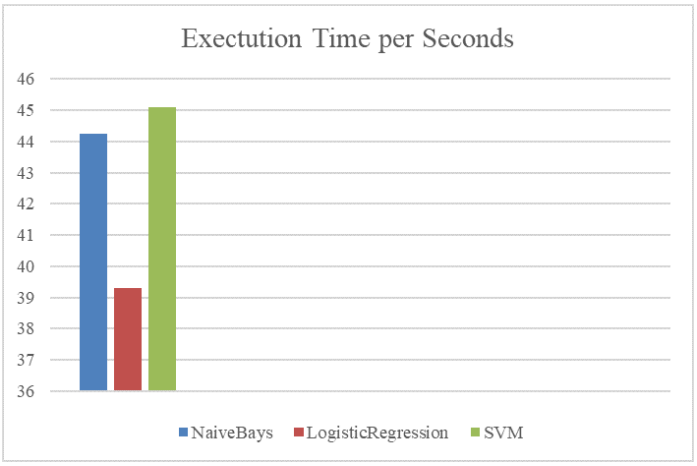


**Figure 5:** Algorithms Execution Time per Seconds

## 4. Conclusion

In this research paper we applied the concept of sentiment analysis for different purpose which is the text embedded inside an image, because pervious research has been done on comments from movies or products and text with image has not been considered. Through our research we concluded that applying supervised method upon text within images has less accuracy in comparison with ordinary comments. In addition support vector machine has higher accuracy than other models, while Logistic Regression model is a faster model than others. This could guild future researchers and application developers to apply support vector machine for better accuracy and Logistic Regression for faster performance.

## References


1. Do, Hai Ha, P. W. C. Prasad, Angelika Maag, and Abeer Alsadoon. "Deep learning for aspect-based sentiment analysis: a comparative review." *Expert Systems with Applications* 118 (2019): 272-299.
2. Abdullah, A."Facial Expression Identification System Using Fisher Linear Discriminant Analysis and K- Nearest Neighbor Methods." *ZANCO Journal of Pure and Applied Sciences*, Vol. 31, no.2, pp. 9-13, doi:10.21271/zjpas.31.2.2. Apr. (2019)
3. Gajarla, V. and Gupta, A, Emotion detection and sentiment analysis of images. "*Georgia Institute of Technology"*. (2015)
4. Choi, K.W., Aich, S. and Kim, H.C., 2018, June. A Machine Learning Approach to Predict Happiness Based on Sentiment Analysis of Twitter Data. "*In international conference on future information & communication engineering"* (Vol. 10, No. 1, pp. 239-241).
5. Amolik, A., Jivane, N., Bhandari, M. and Venkatesan, M.,. Twitter sentiment analysis of movie reviews using machine learning techniques*." International Journal of Engineering and Technology"*, (2016) 7(6), pp.1-7.
6. Gautam, G. and Yadav, D.,Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In 2014 Seventh International Conference on Contemporary Computing (IC3) (pp. 437-442). IEEE. August (2014)
7. Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K. and Taha, K.,. Efficient machine learning for big data: A review. Big Data Research, 2(3), pp.87-93. (2015)
8. Marouli, G.,. Comparison between Maximum Entropy and Naïve Bayes classifiers: Case study; Appliance of Machine Learning Algorithms to an Odesk's Corporation Dataset.( 2014)
9. Schrider, D.R. and Kern, A.D., Supervised machine learning for population genetics: a new paradigm. Trends in Genetics, 34(4), pp.301-312.( 2018)
10. Odena, A., Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv*:1606.01583.(2016)
11. Singh, A., Thakur, N. and Sharma, A., March. A review of supervised machine learning algorithms. In 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1310-1315). *IEEE*. (2016)
12. Mishne G, Glance NS. Predicting movie sales from blogger sentiment. InAAAI spring symposium: computational approaches to analyzing weblogs 2006 Mar 27 (pp. 155-158).
13. R. Wankhede and A. N. Thakare, "Design approach for accuracy in movies reviews using sentiment analysis,"*International conference of Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, (2017), pp. 6-11, doi: 10.1109/ICECA.2017.8203652.
14. Kouloumpis E, Wilson T, Moore J. Twitter sentiment analysis: The good the bad and the omg!. "*InFifth International AAAI conference on weblogs and social media"*. (2011) Jul 5.
15. Bae Y, Lee H. Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. "*Journal of the American Society for Information Science and Technology"*. (2012). Dec;63(12):2521-35.
16. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. "*J Med Internet Res"* (2013);15(11):e239
17. Wei W, Gulla JA. Sentiment learning on product reviews via sentiment ontology tree. "*InProceedings of the 48th Annual Meeting of the Association for Computational Linguistics".* 2010 Jul (pp. 404-413).
18. Jebaseeli AN, Kirubakaran E. A survey on sentiment analysis of (product) reviews. International Journal of Computer Applications. 2012 Jan 1;47(11).
19. Fang X, Zhan J. Sentiment analysis using product review data. Journal of Big Data. 2015 Dec 1;2(1):5.
20. R. K. Bakshi, N. Kaur, R. Kaur and G. Kaur, "Opinion mining and sentiment analysis," *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2016, pp. 452-455.
21. Russell, S.J. and Norvig, P., 2016.Artificial intelligence: a modern approach. Malaysia. Pearson Education Limited. Rycroft-Malone, J. The PARIHS framework—A framework for guiding the implementation of evidenceĞ based practice. "*Journal of nursing care quality"*. 19(4), .(2004) pp.297-304.
22. Goel, S., Madhok, R., & Garg, S. Proposing Contextually Relevant Quotes for Images. Advances in Information Retrieval." 40th European Conference on Information Retreival" 591–597. . (2018) doi:10.1007/978-3-
23. Abd allah, H., W. Yassen, and khalil Alsaif. "Feature Extraction of Images Texture Based on Co-Occurrence Matrix". *ZANCO Journal of Pure and Applied Sciences*, Vol. 31, no. 2, May (2019), pp. 60-69, doi:10.21271/zjpa