# A Semantically Enhanced Deep Neural Network Framework for Reputation System in Web Mining for Covid-19 Twitter Dataset

**Shivani Yadao***
Department of Computer Science and Engineering, Lincoln University College, Selangor, D.E., Malaysia.
E-mail: shivaniyadao123@gmail.com

**Dr.A. Vinaya Babu**
Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women Hyderabad, India.
E-mail: Avb1222@gmail.com

**Dr. Midhunchakkaravarthy Janarthanan**
Department of Computer Science and Multimedia, Lincoln University College, Selangor, D.E., Malaysia.

**Dr. Amiya Bhaumik**
Lincoln University College, Selangor, D.E., Malaysia.

*Received September 21, 2021; Accepted December 17, 2021*
*ISSN: 1735-188X*
*DOI: 10.14704/WEB/V19I1/WEB19258*

## Abstract

With the web containing a huge amount of information, the extraction of application oriented understandable data has become easier with web mining. Web Mining is the area that is derived from data mining. Unlike data mining, web mining is used to extract interesting patterns from the information available on the web. When used with deep learning, the pattern recognition becomes much easier. Deep learning works in the same way, how a human brain works in terms of predicting the outcomes when a bulk of information is provided. It deals with mathematical models that recognize the patterns efficiently. The different types of web mining techniques, namely: web content mining (WCM), web structure mining (WSM) and web usage mining (WUM) persists. Researchers and economists around the globe are keen in knowing the impact of the pandemic on the society's economic status; this work helps find the same using reputation system. As twitter is a hub of different opinions of public, we work with covid- 19 data set from twitter. A reputation system helps finding the socio economic status of the tweets regarding covid-19 dataset. This paper has proposed a framework in which the web mining is implemented using a semantic enhanced deep neural network technique for the reputation system.

## Keywords

Web Content Mining (WCM), Web Usage Mining (WUM), Web Structure Mining (WSM), Semantic Enhanced Framework, Deep Neural Networks, Reputation System.

## Introduction

The term Reputation System refers to understanding the level of prominence that a website or an application has, depending upon the opinion of its user's worldwide. For any business to attain a good amount of success, especially ones that are online, it's important to check on its reliability in terms of its user reviews. Online reputation systems persist important considering developing effective and dependable online services in a variety of businesses, including e-commerce and e-learning, as well as data analysis considering socioeconomic consequences in research. However, due to a lack of mining optimization efficiency, defects and limitations in online reputation systems have been widely recognized. Users who persist dissatisfied among system resolve, soon stop using it. Many authors and papers claim that existing online reputation systems have key problems, resulting in client loss and putting e-commerce and e-learning services at danger regarding significant revenue loss due to increased number regarding errors and shortcomings in web data models. Some of the existing algorithms use traditional machine learning algorithms for classification of such opinion mining. However, what makes this paper interesting and unique is its application of deep neural networks where the machine learns itself with least human intervention. Not only does it use deep learning, but also merges it with the concept of web mining making it different from the existing model. This paper has proposed a model in which the reputation system for covid-19 twitter data sets is found.

In today's time when the pandemic has affected the world, the growth in technology has drastically improved and people have switched to internet for almost every day to day activity. Almost every area including business, medical field, educational field, banking field, office meetings, etc are using resources available online. An immense amount of information that is there is made available on web. This makes web mining an important field of research. A derived field of data mining is web mining. Mining means to find important client required data in the form of a pattern. Were data mining deals with mining these patterns from a large database, web mining deals with the same mining of data but from the information available on the web (V. Medvedev, et al., 2017); (S. M. Huded, et al., 2019); (S. Taha Ahmed, et al., 2018). There are three categories of web mining namely; Web Content Mining (WCM), Web Structure Mining (WSM), Web Usage Mining (WUM) (T. A. Al-asadi, et al., 2017); (M. J. H. Mughal, 2018); (C. Li,

2021). In Web Content Mining the data that is available on web pages in the form of text, images, videos etc is mined (R. H. Salman, et al., 2020); (E. T. John, et al., 2016). A website that is available on a web or application server is a collection of interrelated web pages. Web Structure Mining finds the relation between these pages so as to get the information of their structure (N. Pradhan and V. Dhaka, 2020). Web Usage Mining on other hand deals with the utility of a web user or client, making it easy to find the common searches and common area of interest of a user (S. Shanthi, 2017); (Bhoumik, S., et al., 2020); (Chhabra, H., et al., 2020).

There are some common steps of mining which are used by all the three web mining techniques. Figure 1 shows the subtasks of the mining techniques.

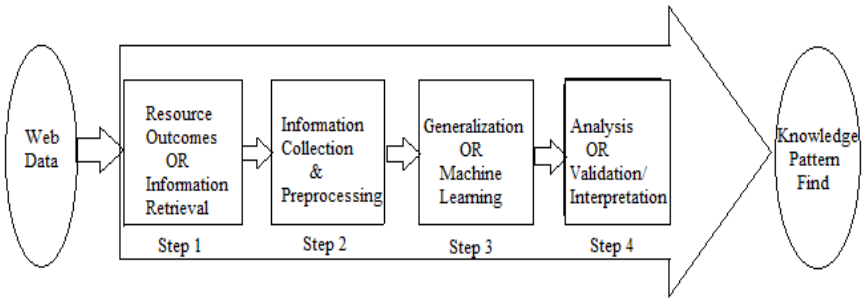



**Figure 1 Subtasks regarding Web Mining**

Step 1: Resource Outcomes – This is preliminary subtask for which data is extracted or guidance retrieval for web is done.
Step 2: Preprocessing - during this step, data collected is preprocessed. Preprocessing is nothing but removing unnecessary data that creates error during further steps.
Step 3: Generalization - This step deals with bringing data to a general easy, understandable and modifiable form.
Step 4: Analysis - This is the most important step which contains logic regarding how exactly data is supposed to be converted into client defined required patterns. End result is knowledge pattern.

There exists a big variation regarding algorithms and software's that are used among these three different types of web mining. PageRank, weighted PageRank, topic sensitive PageRank, Hits, distance rank, SimRank, and other web structure mining techniques exist. Most popular structure mining methods exist PageRank and SimRank. PageRank is most fundamental and widely used algorithm, but SimRank is a newer algorithm that may occur using multiple techniques that achieve better results. PageGather, CDL4, Leader, Cobweb, Iterate, are the other web usage mining algorithms. Web content mining

algorithms include correlation algorithm considering relevance ranking, Weighted Page Content Rank, Cluster hierarchy, fuzzy c-mean algorithm, and construction algorithm (Mohd Shoaib and Ashish K. Maurya, 2014); (Joseph, F. J. J., & Auwatanamongkol, S., 2016).

## Deep Learning with Web Mining

Artificial neural networks are used during deep learning that execute complex computations on enormous volumes of data. It's a form of machine learning that's based on human brain's structure and function (Biran, O., and Cotton, C. 2017). Machines are trained using in-depth reading algorithms of deep learning that learn from the examples. Deep learning is extensively used among industries such as health care, ecommerce, entertainment, and advertising. Applications in deep learning are shown in Figure 2 (Cao, S., Lu, W., and Xu, Q. 2016); (Mohan, V., et al., 2019).

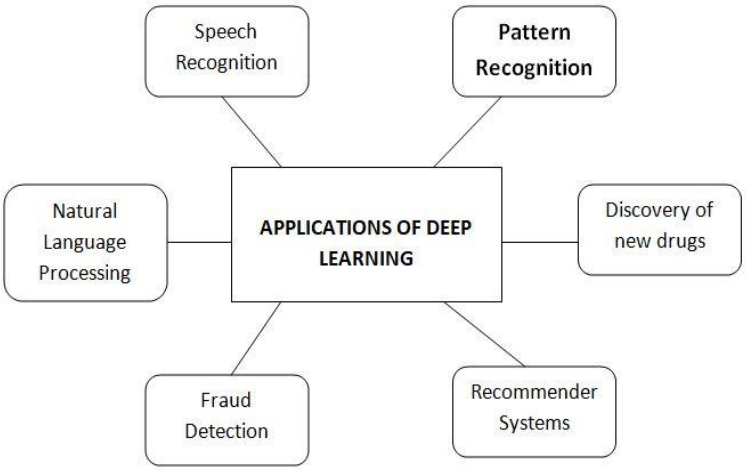



**Figure 2 Applications in Deep Learning**

As it is observed from figure 2, one amongst many applications in deep learning is pattern recognition which is also the main task under web mining. Idea regarding combining two domains ‘Web mining’ and ‘deep learning’ comes from here. This paper has sentiment analysis as a major part of the process where deep learning is considered for reputation system (Dai, J., Wang, Y., Qiu, X., Ding, D., Zhang, Y., Wang, Y., et al. 2018). Computational examination of people's opinions, sentiments, attitudes, and emotions conveyed using written language is known as sentiment analysis. Sentiment analysis creates a digital version regarding an opinionated text. Sentiment classification is a method for analyzing subjective data within a text and then extracting an opinion.

Sentiment analysis process is about extracting information from people's opinions, assessments, and feelings about entities, events, and their properties. A sentence's polarity parameter is classified into three categories: positive, negative, and neutral. In Contact or call centers and social media monitoring, sentiment analysis is commonly employed. Sentiment returns a tuple regarding form (polarity, subjectivity), where polarity represents a float of range [-1.0, 1], and subjectivity a float of range [0.0, 1.0]. In subjectivity, where the range is [0.0, 1.0], 0.0 represents very objective and 1.0 represents very subjective, whereas in polarity with range [-1.0, 1], -1.0 is very negative and 1.0 is very positive.

## Existing System

Semantic enhanced deep neural network framework is compared and builds over the paper that uses Support Vector Machine algorithm for Twitter Sentiment Recognition (V Uday Kumar, et al., 2019). These methods frequently rely on supervised classification algorithms, in which sentiment detection is defined as positive or negative binary items. To train classifiers, this method requires labeled data. During this approach, it becomes clear that negative (e.g., not beautiful) and intensifying components regarding a word's local context must be considered (e.g. Very beautiful). However, a basic paradigm considering the creation of a feature Vector is demonstrated:

1. Tag each tweet among a part regarding speech tagger.
2. Gather every adjectives considering entire Twitter post.
3. Create a popular word set using top N adjectives.
4. Go through every respective tweet during experimental set to find one you want. Make following list:

    - Presence, absence, or frequency regarding each term.
    - Number regarding good words.
    - Number regarding negative words.

Though, one of the important advantages of this system is its capability of generating and adjusting the trained data models with reference to the specified need or requirements, it has a research gap in terms of the labeled data. Having only the labeled data is a costly affair and also not adaptable for all the systems.

The experiment and analysis for this paper was done using the general twitter data set. However the accuracy of the system is less as there is no module to have a semantic bag of word model referring to multiple meanings of a word.

## Semantic Enhanced Weighting Deep Neural Network Framework for Reputation Systems

A large number of data has to be defined and understood in order to design and implement a good website. For improving system efficiency, adoption of twitter data ontology learning model used among a Semantic web mining-based Deep Neural Network (DNN) framework considering socio economic impact of covid-19. This work proposes a content management meta-model that combines web's usability among semantic web's expressivity and flexibility (Guo, Y., Liu, Y., Georgiou, T. et al., 2018); (Shimoda W, Yanai K, 2016); (He K, Zhang X, Ren S, Sun J, 2016). Figure 3 shows basic architecture regarding semantic enhanced weighting deep neural network framework (Jin B, Ortiz-Segovia MV, Süsstrunk S, 2017). The new contributions in the paper are the modules: Semantic item generation, where the words having multiple but similar meanings are collected and processed using deep learning; and the other module where the algorithm for SWEDNNF is given (Khoreva A, Benenson R, Hosang J, Hein M, Schiele B, 2017); (Lin G, Shen C, Van Den Hengel A, Reid I, 2017).

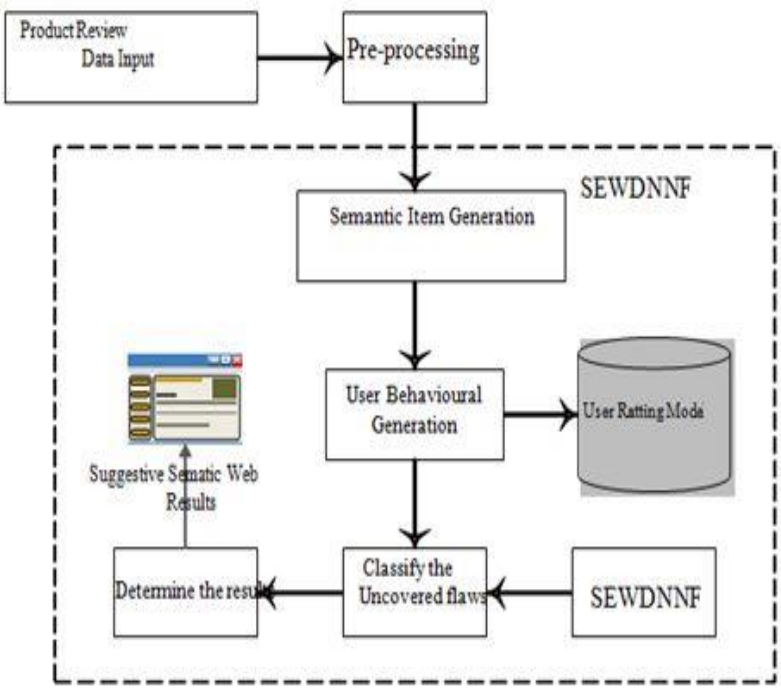
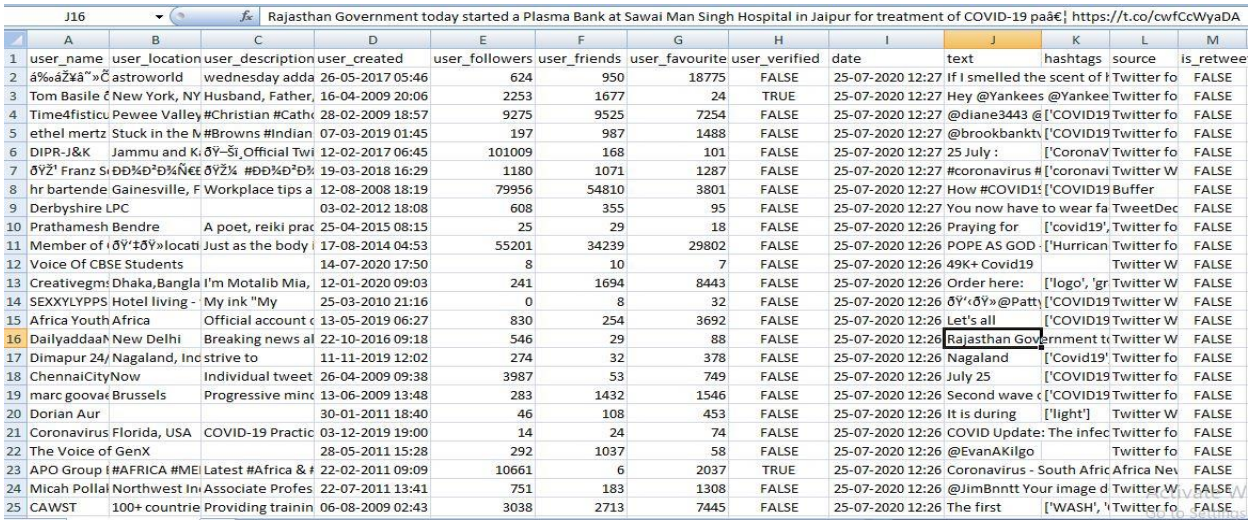



**Figure 3 SEWDNNF Model**

The research's major goal is to make web resources more machine-accessible and manipulable at providing metadata that characterises Web content in a user requirement specific data format. We will examine several semantic web based modules on covid-19 twitter data set.

The covid-19 twitter data set is collected from 3 different sources namely; Kaggle (80%), direct social media (twitter 15%), web crawling (websites /online news reports 5%). The data set has a total of 179000 entries of tweets that were done in the year 2020 when the pandemic was on its peak and later lowered down. The total number of columns is 13 as shown in the figure 4. The fields include "user name, user location, user description, user created, user followers, user friends, user favourite, user verified, date, text, hash tags, source, is re-tweeted".



J16 — Rajasthan Government today started a Plasma Bank at Sawai Man Singh Hospital in Jaipur for treatment of COVID-19 paâ€¦ https://t.co/cwfCcWyaDA

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | user_name | user_location | user_description | user_created | user_followers | user_friends | user_favourite | user_verified | date | text | hashtags | source | is_retwee |
| 2 | á‰¤åŽ¥â˜»Č | astroworld | wednesday adda | 26-05-2017 05:46 | 624 | 950 | 18775 | FALSE | 25-07-2020 12:27 | If I smelled the scent of h | | Twitter fo | FALSE |
| 3 | Tom Basile ć | New York, NY | Husband, Father, | 16-04-2009 20:06 | 2253 | 1677 | 24 | TRUE | 25-07-2020 12:27 | Hey @Yankees @Yankee | | Twitter fo | FALSE |
| 4 | Time4fisticu | Pewee Valley | #Christian #Cath | 28-02-2009 18:57 | 9275 | 9525 | 7254 | FALSE | 25-07-2020 12:27 | @diane3443 @ | ['COVID19 | Twitter fo | FALSE |
| 5 | ethel mertz | Stuck in the M | #Browns #Indian | 07-03-2019 01:45 | 197 | 987 | 1488 | FALSE | 25-07-2020 12:27 | @brookbanktv | ['COVID19 | Twitter fo | FALSE |
| 6 | DIPR-J&K | Jammu and K | ðŸ–Ši Official Twi | 12-02-2017 06:45 | 101009 | 168 | 101 | FALSE | 25-07-2020 12:27 | 25 July : | ['CoronaV | Twitter fo | FALSE |
| 7 | ðŸŽ¹ Franz S | ÐĐ¾Ð³Ð¾Ñ€ | ðŸŽ¼ #ÐĐ¾Ð³Ð¾ | 19-03-2018 16:29 | 1180 | 1071 | 1287 | FALSE | 25-07-2020 12:27 | #coronavirus # | ['coronavi | Twitter W | FALSE |
| 8 | hr bartende | Gainesville, F | Workplace tips a | 12-08-2008 18:19 | 79956 | 54810 | 3801 | FALSE | 25-07-2020 12:27 | How #COVID19 | ['COVID19 | Buffer | FALSE |
| 9 | Derbyshire | LPC | | 03-02-2012 18:08 | 608 | 355 | 95 | FALSE | 25-07-2020 12:27 | You now have to wear fa | | TweetDec | FALSE |
| 10 | Prathamesh | Bendre | A poet, reiki prac | 25-04-2015 08:15 | 25 | 29 | 18 | FALSE | 25-07-2020 12:26 | Praying for | ['covid19', | Twitter fo | FALSE |
| 11 | Member of | ðŸ‘‡ðŸ»locati | Just as the body | 17-08-2014 04:53 | 55201 | 34239 | 29802 | FALSE | 25-07-2020 12:26 | POPE AS GOD | ['Hurrican | Twitter fo | FALSE |
| 12 | Voice Of CBSE | Students | | 14-07-2020 17:50 | 8 | 10 | 7 | FALSE | 25-07-2020 12:26 | 49K+ Covid19 | | Twitter W | FALSE |
| 13 | Creativegm | Dhaka,Bangla | I'm Motalib Mia, | 12-01-2020 09:03 | 241 | 1694 | 8443 | FALSE | 25-07-2020 12:26 | Order here: | ['logo', 'gr | Twitter W | FALSE |
| 14 | SEXXYLYPPS | Hotel living - | My ink "My | 25-03-2010 21:16 | 0 | 8 | 32 | FALSE | 25-07-2020 12:26 | ðŸ‘‹ðŸ»@Patt | ['COVID19 | Twitter W | FALSE |
| 15 | Africa Youth | Africa | Official account | 13-05-2019 06:27 | 830 | 254 | 3692 | FALSE | 25-07-2020 12:26 | Let's all | ['COVID19 | Twitter W | FALSE |
| 16 | Dailyaddaa | New Delhi | Breaking news al | 22-10-2016 09:18 | 546 | 29 | 88 | FALSE | 25-07-2020 12:26 | Rajasthan Gov | rnment t | Twitter W | FALSE |
| 17 | Dimapur 24/ | Nagaland, Ind | strive to | 11-11-2019 12:02 | 274 | 32 | 378 | FALSE | 25-07-2020 12:26 | Nagaland | ['Covid19' | Twitter fo | FALSE |
| 18 | ChennaiCity | Now | Individual tweet | 26-04-2009 09:38 | 3987 | 53 | 749 | FALSE | 25-07-2020 12:26 | July 25 | ['COVID19 | Twitter fo | FALSE |
| 19 | marc goova | Brussels | Progressive minc | 13-06-2009 13:48 | 283 | 1432 | 1546 | FALSE | 25-07-2020 12:26 | Second wave | ['COVID19 | Twitter fo | FALSE |
| 20 | Dorian Aur | | | 30-01-2011 18:40 | 46 | 108 | 453 | FALSE | 25-07-2020 12:26 | It is during | ['light'] | Twitter W | FALSE |
| 21 | Coronavirus | Florida, USA | COVID-19 Practic | 03-12-2019 19:00 | 14 | 24 | 74 | FALSE | 25-07-2020 12:26 | COVID Update: The infec | | Twitter fo | FALSE |
| 22 | The Voice of | GenX | | 28-05-2011 15:28 | 292 | 1037 | 58 | FALSE | 25-07-2020 12:26 | @EvanAKilgo | | Twitter fo | FALSE |
| 23 | APO Group | #AFRICA #ME | Latest #Africa & | 22-02-2011 09:09 | 10661 | 6 | 2037 | TRUE | 25-07-2020 12:26 | Coronavirus - South Afric | Africa Nev | | FALSE |
| 24 | Micah Pollal | Northwest In | Associate Profes | 22-07-2011 13:41 | 751 | 183 | 1308 | FALSE | 25-07-2020 12:26 | @JimBnntt Your image d | | Twitter W | FALSE |
| 25 | CAWST | 100+ countrie | Providing trainin | 06-08-2009 02:43 | 3038 | 2713 | 7445 | FALSE | 25-07-2020 12:26 | The first | ['WASH', ' | Twitter fo | FALSE |

**Figure 4 Covid-19 Dataset entries**

## Experimental Model

The Semantic enhanced weighting deep neural network framework designed among below modules.

### 1. Pre-processing

After mining the appropriate data set which is in this case Covid-19 twitter data, the tweets retrieved is fetch during first stage. This stage involves preprocessing data, which involves removing unidentifiable or unneeded information such as content timestamps, embedded links, videos, unreadable data etc. Such outputs are usually unimportant and may cause our system to produce erroneous results.

### 2. Semantic Item Generation

Semantic means to have different meanings regarding a single word, i.e. either logical or language oriented. We create a pool of such semantics so as to generate multiple meanings regarding an adjective to check its polarity (Liu Y, Guo Y, Lew MS, 2017);

(Liu S, Qi X, Shi J, Zhang H, Jia J, 2016); (Lin G, Shen C, van den Hengel A, Reid I, 2016); (Saleh F, et al., 2016). For example the term “destination” has the same meaning as “last stop”. However these different shades of terms are used by users in accordance with their applicability. In this step we also perform deep learning in the form of text categorization. This is done so as to extract certain adjectives/terms and there recurring occurrence from the data set. This occurrence will ultimately tell us the intensity or strength of that word in the statement helping us find the polarity after checking its semantic from the pool of semantic words.

For text categorization, deep neural network makes the use of 1D structure regarding text data through convolution layers. While considering text categorization, Deep neural network requires vector representation regarding data that preserves internal locations as input. Each word would occur treated as a pixel, and each document will treat as an image regarding |D|*1 pixel among |V| channels, where V is vocabulary regarding documents |D|. Each convolution layer consists of region vector that contains pixels i.e. words. Output in convolution layer consists the variable sized output. This variable sized output was passed via pooling layer to produce a fixed size output. Pooling method applied in text is typically max-pooling over entire data.

For illustration consider a text document that has been categorized using a Deep neural network. Each document D has words which exist considered as pixels. Among provided documents, vocabulary |V| was created. Each document was considered as an image among |D|*1 pixels among |V| channels. This illustration considers two documents D1 and D2 among provided vocabulary V. Now we will create a region vector which will basically check whether inside the vocabulary, the word from chosen document is present or not. If the word is present then write 1, if not present then write 0. Region vectors considering input document D1and D2 was as shown below:

Input text regarding D1 = {this is good}
Input text regarding D2 = {Doing good work}
Vocabulary: {This is doing good work}

| | | |
|---|---|---|
| Region vector1: | 1 | This |
| | 1 | is |
| | 0 | doing |
| | 1 | good |
| | 0 | work |

| | | |
|---|---|---|
| | 0 | T *h*is |
| Region vector 2: | 0 | is |
| | 1 | doing |
| | 1 | good |
| | 1 | work |

The utility of creating the vocabulary and region vector is that when we merge certain tweets from different users to create a common vocabulary, we can analyze how often a word is used in it and later check for its semantics to build a strong opinion mining system.

## 3. User Behavioral Generation / User Rating Mode

This stage gets input from Semantic item generation and fetches it here among user information. This data is saved and retrieved from within a database of User Rating mode.

During User behavioral generation, we train the system using logistic regression. The reason of selecting this classification is that checking the polarity is basically either positive or negative normally, and logistic regression works well with binary input variables. Also the table 3, of comparison between 3 different classifiers indicates that logistic regression works better amongst the three. Input document or sentence is classified into two category or three category classification. When response variable is binary, often used model is linear logistic regression model. Response variable takes values as {0, 1}. Considering a brief description, consider a given dataset x containing n records. Each record consists of features or predictor variables or attributes and a binary outcome variable or class or response. This binary outcome variable preserve assume only two possible values 0 or 1 i.e. either positive or negative. Goal regarding logistic regression is, using the provided dataset to create a predictive model regarding outcome variable. Logistic function σ(t)is defined as follows:

$$y = f(x) \qquad (4.1)$$

$$Y = \sigma(t) = \left( \frac{1}{1+e^{-t}} \right)$$

t– Linear function regarding a single explanatory variable x

$$P(y|x,\Theta) = \sigma(\theta^T x) = \left( \frac{1}{1+e^{-\theta x}} \right)$$

$$Y \in \{0,1\}$$

x – Feature Vector

x $\in\Theta\in R^N$ -N dimensional feature vector

Θ - Parameters of the logistic regression

## 4. Classify Uncovered Flaws

The statements or reviews generated since previous stage often contain blanks or null values. This stage removes these uncovered flaws and adds logic regarding SEWDNNF via it so as to generate Semantic web results in the form of a positive polarity or negative polarity.

## 5. SEWDNNF

In this step algorithm for semantic enhanced weighting deep neural network framework is given. Second phase about our proposed method begins among us attempting to determine polarity about text in question. If emoticons persist in statements, they resolve and are used to compute statement's overall polarity. We're looking considering sentences where polarity detection isn't quite right or when stated sentiment isn't expressed quite right. We also try separating opinion terms in sentence in connection with given sentence's concepts.

a. We train the system to recognize the relationship between words in various contexts. We check the number of times the word has occurred in the statement so as to help find how strong the statement is negative or positive.
b. Once the opinion words are identified with Context, we can find the polarities of the words. Once opinion words have been found using Context, polarities about words can be determined.
c. To aid in detection about concepts involved, we use a huge dataset that expresses a wide range about complicated and ambiguous emotions via train our system. This Data presented via system in an unsupervised manner.

Response regarding default training dataset falls into one of two categories; either positive or negative. We also consider the neutral parameter after we first identify positive and negative. Hence, during this paper, training dataset classified into multiple levels using qualitative prediction. Most widely used classifier that is used in predicting a qualitative response is logistic regression. Logistic regression model has been a widely used model when categorical response is expected. Predicted probabilities outcome regarding training dataset lies between 0 and 1. As a result, binomial distribution is also employed. Accuracy regarding model performance on trained data of unigram and bigram models is done using Term Frequency-Inverse Document Frequency (TF-IDF) transformation technique via normalize Document-Term Matrix (DTM) as shown in figure 5.

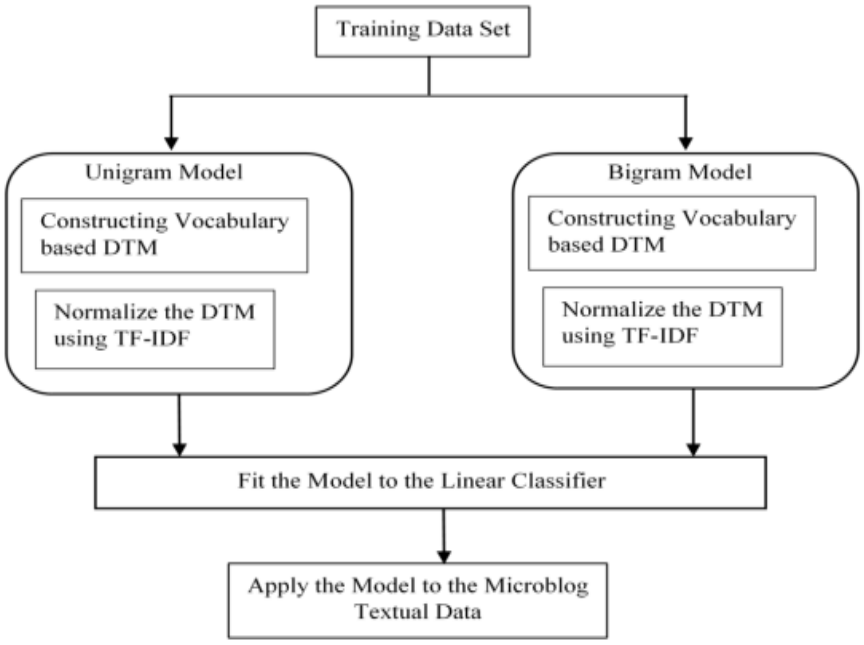



**Figure 5 Process to Perform Semantic Analysis**

TF-IDF, short form of "term frequency–inverse document frequency", is a mathematical measure considering deciding how significant a word through an archive in an assortment or document or corpus is. It's broadly utilized as a weighting factor in look thinking about data recovery, text mining, and client demonstrating. The value of TF-IDF is directly proportional to the amount of occurrence of the word, in terms of the frequency of recurring words in a document. Related with Term-weighting approach, TF-IDF stands one amongst the most broadly used scheme today.

A document-term matrix is a numerical lattice that shows recurrence about terms in an assortment about reports or documents. Like any other matrix, this matrix too is a combination of rows and columns, where the rows will have number of reports or documents and columns will have entries of terms. The rows and columns are shuffled in case of transpose or term document matrix.

While making an informational index about terms that show up in a corpus about archives, the document-term matrix has terms in the columns and documents in the rows. Each xy cells addresses number of times word y shows up in report x. As an outcome, each line addressed aside a vector about term counts that addresses content with regards to the record comparing through means of that column. Think about two (short) reports for instance:

- R1 = "I am research scholar"
- R2 = "I am research guide",

Then the document-term matrix would be:

**Table 1 DTM**

| | **I** | **am** | **research** | **scholar** | **guide** |
|---|---|---|---|---|---|
| R1 | 1 | 1 | 1 | 1 | 0 |
| R2 | 1 | 1 | 1 | 0 | 1 |

While raw count about a word normally is the cell value, however there persist numerous ways considering balancing raw counts, including row normalization (i.e. “relative frequency/proportions”) and tf-idf. On each side about term, whitespace or punctuation frequently are used via separate single words (a.k.a. unigrams). This often is referred via as a "bag of words" representation because individual word counts persist retained but not order about words in document.

## Algorithm 1: To Perform Semantic Analysis

Input: Extracted Micro-blogging data (M)
Output: Semantic regarding Micro-blogging data = {0,1} Begin

1. Import training data set among positive or negative Semantic
2. Create DTM, $T_D$ in the training dataset
3. Obtain Normalized DTM, $T_{DN}$ by applying normalization technique TF-IDF transformation on $T_D$
4. Fit Function Regression R applied via $T_{DN}$ for obtaining $RT_{DN}$
5. Construct Micro-blogging textual data M and fit it in $RT_{DN}$ for obtaining $RM_{DN}$
6. Return
   Semantic
   End

## Algorithm 2: Dummy Variable Approach

Input: Predicted Semantic Rate regarding Micro-blogging Textual Data S ($RM_{DN}$) = [0, 1]
Output: Semantic regarding Micro-blogging Textual data = {0, 1}
Begin

1. Import predicted Semantic rate regarding Micro-blogging Textual data in the $RM_{DN}$ results S(RMDN)
2. If $S(RM_{DN}) > 0.5$ then

a. S = 1

3. Else
   a. S = 0
4. Endif
5. Return S
6. End

As logistic regression typically is used among a qualitative response, so response regarding result should occur one amongst the categories: 0 or 1, but predicted probability regarding micro-blogging textual unseen data lies between 0 and 1. Predicted probabilities outcome will occur ordered as 0 till 0.35 as negative Semantic rate, 0.35 till 0.65 ordered as neutral Semantic rate and 0.65 till 1 ordered as positive Semantic rate. Gap between negative and neutral is observed to occur similar with gap between neutral and positive.

To find accuracy, training dataset response should have two possible outcomes. For getting a binary qualitative response, dummy variable approach could potentially be used via code response as follows.

Response = 0; if Predicted micro-blogging text Semantic rate is<0.5
Response = 1; if Predicted micro-blogging text Semantic rate is> 0.5.

To apply linear classifier model via unseen data, a dummy variable created, which takes value 1 considering response if predicted micro-blogging text Semantic rate greater than 0.5 and it takes value 0 considering response if predicted micro-blogging text Semantic rate less than 0.5.

## Performance Metrics

Sorting of public opinion is done by classification of opinions into positive, negative and neutral, thus it becomes the classification problem. (Sherin Mariam John, K. Kartheeban, 2019). For calculation of accuracy, comparison between some selective classifiers such as logistic regression, SVM and decision tree are used on dataset of covid-19 tweets. We calculate the confusion matrix so that we can calculate the other performance metrics of classifier, such as precision, recall and F measure.

$$\text{Accuracy} = \frac{\text{Count of accurately classified field}}{\text{Total Count in the record}}$$

Since the accuracy in above formula is not sufficient in terms of correctness, the formula is updated to the below using the confusion matrix.

**Table 2 Confusion matrix**

| **Actuals** | **Predicted Class** | |
|---|---|---|
| | **Positive Forecasts** | **Negative Forecasts** |
| Actual Positive | TP (Real Positives) | FN (Wrong Negatives) |
| Actual Negative | FP (Wrong Positives) | TN (Real Negatives) |

Accuracy = (TP + TN) / (TP + TN + FP +FN)
Precision (measuring exactness) = TP / (TP + FP)
Recall (measuring completeness) = TP / (TP + FN)
F1 = (2* Precision* Recall) / (Precision + Recall)

**Table 3 Comparison between different classifiers in terms of performance metrics**

| **Classifier** | **Precision** | **Recall** | **F Score** |
|---|---|---|---|
| Logistic Regression | 0.94 | 0.99 | 0.96 |
| SVM | 0.92 | 0.98 | 0.92 |
| Decision Tree | 0.90 | 0.91 | 0.93 |

From the analysis of table 3, Logistic regression was the best classifier method for the covid-19 twitter data set.

**Table 4 Comparison between existing and proposed system in terms of accuracy**

| **Research Model** | **Accuracy** |
|---|---|
| SEWDNNF | 96.80 |
| SVM | 95.70 |

From table 4 it is also seen that when compared to an existing system of classic Support Vector Machine algorithm for Twitter Sentiment Recognition (V Uday Kumar, et al., 2019), the accuracy of proposed model, SEWDNNF is more.

## Conclusion and Future Scope

The figure 6 of analyzed twitter data shows that the opinions of the people on pandemic covid-19 by the mid year were mostly neutral and positive. When compared with older data, these results state an improvement in peoples mind set towards fighting the pandemic and thus social mind set can be predicted as neutrally positive. However from figure 7 which depicts the most frequently occurring words that are observed are panic, crisis and scam. This indicates that the public opinion on economic status of the world isn't very good. The proposed model not just merges two domains, namely: Web Mining

and Deep Neural Networks but also proposes a new model considering online Reputation systems providing a semantic approach in getting polarities regarding covid-19 tweets or reviews that has negative, positive and neutral outputs as shown in figure 6. This paper has opinionated different parameters of the covid-19 dataset to mine a socio economic impact on public sentiments (Adeena Nasir, et al., 2021).

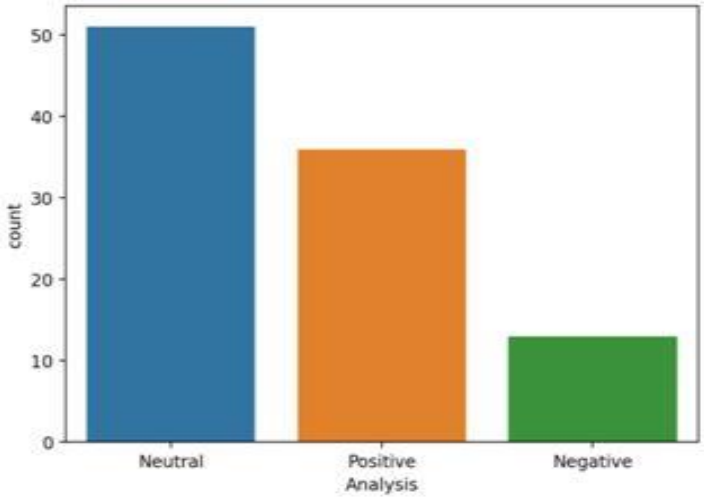



**Figure 6 Analysis of covid-19 tweets in terms of polarity**

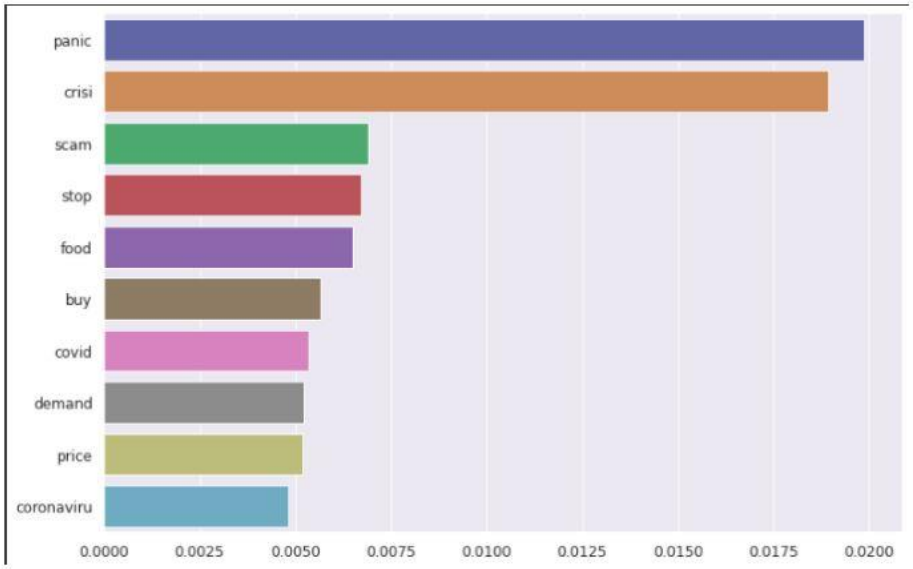



**Figure 7 Most frequently used words**

The Future scope regarding this research work can be network security. As web is a collection containing sensitive data, there exist various techniques available for retrieving this information directly or indirectly. So security during web mining is an important aspect considering further researches.

## References


Medvedev, V., Kurasova, O., Bernatavičienė, J., Treigys, P., Marcinkevičius, V., & Dzemyda, G. (2017). A new web-based solution for modelling data mining processes. *Simulation Modelling Practice and Theory*, *76*, 34-46.

Huded, S.M., Balutagi, S., & Ranjan, A. (2019). Mapping of literature on data mining by j-gate database.

Taha Ahmed, S., Al-hamdani, R., & Crook, M. (2018). Studying regarding Educational Data Mining Techniques. *International Journal regarding Advanced Research during Science, Engineering and Technology*, *5*, 5742-5750,

Al-asadi, T.A., Obaid, A.J., Hidayat, R., & Ramli, A.A. (2017). A survey on web mining techniques and applications. *International Journal on Advanced Science Engineering and Information Technology*, *7*(4), 1178-1184.

Mughal, M.J.H. (2018). Data mining: Web data mining techniques, tools and algorithms: An overview. *Information Retrieval*, *9*(6).

Li, C. (2021). Research on an Enhanced Web Information Processing Technology based on AIS Text Mining. *Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering)*, *14*(1), 29-36.

Salman, R.H., Zaki, M., & Shiltag, N.A. (2020). A Studying regarding Web Content Mining Tools. *Al-Qadisiyah Journal regarding Pure Science*, *25*, 1-16.

John, E.T., Skaria, B., & Shajan, P. (2016). An Overview regarding Web Content Mining Tools. *Bonfring International Journal regarding Data Mining*, *6*, 01-03.

Pradhan, N., & Dhaka, V.S. (2020). Comparison-based study of page rank algorithm using web structure mining and web content mining. In *Smart Systems and IoT: Innovations in Computing*, Springer, 719-729.

Shanthi, S. (2017). Survey on web usage mining using association rule mining. *International Journal of Innovative Computer Science & Engineering*, *4*(3), 65-67.

Shoaib, M., & Maurya, A.K. (2014). Comparative Study of Different Web Mining Algorithms to Discover Knowledge on the Web. *In Proceedings of Elsevier Second International Conference on Emerging Research in Computing, Information, Communication and Application (ERCICA-2014)*, *3*, 648-654.

Uday, K.V., Zeelan, B.C.M.A.K., Vikas, C.M., Sai, M.D., & Anish, K. (2019). Twitter Sentiment Recognition using Support Vector Machine. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(4), 2019.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. *In IJCAI-17 workshop on explainable AI (XAI)*, 8(1), 8-13.

Cao, S., Lu, W., & Xu, Q. (2016). Deep neural networks for learning graph representations. *In Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1).

Dai, J.J., Wang, Y., Qiu, X., Ding, D., Zhang, Y., Wang, Y., & Song, G. (2019). Bigdl: A distributed deep learning framework for big data. *In Proceedings of the ACM Symposium on Cloud Computing*, 50-60.

Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, *7*(2), 87-93. https://doi.org/10.1007/s13735-017-0141-z

Shimoda, W., & Yanai, K. (2016). Distinct class-specific saliency maps for weakly supervised semantic segmentation. *In European Conference on Computer Vision*, Springer, 218-234.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

Jin, B., Ortiz Segovia, M.V., & Susstrunk, S. (2017). Webly supervised semantic segmentation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3626-3635.

Khoreva, A., Benenson, R., Hosang, J., Hein, M., & Schiele, B. (2017). Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 876-885.

Lin, G., Shen, C., Van Den Hengel, A., & Reid, I. (2017). Exploring context with deep structured models for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1352-1366.

Liu, Y., Guo, Y., & Lew, M.S. (2017). On the exploration of convolutional fusion networks for visual recognition. *In International conference on multimedia modeling*, Springer, 277-289.

Liu, S., Qi, X., Shi, J., Zhang, H., & Jia, J. (2016). Multi-scale patch aggregation (mpa) for simultaneous detection and segmentation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3141-3149.

Lin, G., Shen, C., Van Den Hengel, A., & Reid, I. (2016). Efficient piecewise training of deep structured models for semantic segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 3194-3203.

Saleh, F., Aliakbarian, M.S., Salzmann, M., Petersson, L., Gould, S., & Alvarez, J.M. (2016). Built-in foreground/background prior for weakly-supervised semantic segmentation. *In European conference on computer vision*, Springer, 413-432.

John, S.M., & Kartheeban, K. (2019). Sentiment Scoring and Performance Metrics Examination of Various Supervised Classifiers. *International Journal of Innovative Technology and Exploring Engineering*, *9*(2S2).

Bhoumik, S., Chatterjee, S., Sarkar, A., Kumar, A., & John Joseph, F.J. (2020). Covid 19 Prediction from X Ray Images Using Fully Connected Convolutional Neural Network. *In CSBio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*, 106-107.

Joseph, F.J.J., & Auwatanamongkol, S. (2016). A crowding multi-objective genetic algorithm for image parsing. *Neural Computing and Applications, 27*(8), 2217-2227.

Mohan, V., Chhabra, H., Rani, A., & Singh, V. (2019). An expert 2DOF fractional order fuzzy PID controller for nonlinear systems. *Neural Computing and Applications, 31*(8), 4253-4270.

Chhabra, H., Mohan, V., Rani, A., & Singh, V. (2020). Robust nonlinear fractional order fuzzy PD plus fuzzy I controller applied to robotic manipulator. *Neural Computing and Applications, 32*(7), 2055-2079. https://doi.org/10.1007/s00521-019-04074-3

Nasir, A., Shah, M.A., Ashraf, U., Khan, A., & Jeon, G. (2021). An intelligent framework to predict socioeconomic impacts of COVID-19 and public sentiments. *Computers & Electrical Engineering, 96*.