Original Article

A Combined Approach of Web Content Mining and Neural Networks for Predicting the societal impact of covid-19 through twitter

Shivani Yadao[1], Dr. A. Vinaya Babu[2], Dr. Midhunchakkaravarthy Janarthanan[3], Dr. Amiya Bhaumik[4]

[1]Dept. of Computer Science and Engineering, Lincoln University College, Selangor, D.E., Malaysia

[2]Dept of Computer Science and Engineering, Stanley College of Engineering and Technology for Women Hyderabad, India

[3]Dept of Computer Science and Multimedia, Lincoln University College, Selangor, D.E., Malaysia

[4]Lincoln University College, Selangor, D.E., Malaysia

Email: shivaniyadao123@gmail.com

DOI: 10.47750/pnr.2022.13.S06.008

Abstract

Web data mining became an easy and important platform for retrieval of useful information. Users prefer World Wide Web more to upload and download data. As increasing growth of data over the internet, it is getting difficult and time consuming for discovering informative knowledge and patterns. Digging knowledgeable and user queried information from unstructured and inconsistent data over the web is not an easy task to perform. Different mining techniques are used to fetch relevant information from web (hyperlinks, contents, web usage logs). Web data mining is a sub discipline of data mining which mainly deals with web. Web data mining is divided into three different types: web structure, web content and web usage mining. Page content mining also called web scrapping is the technique under web mining that is used for extracting the web data. It is used in combination with the deep neural network algorithm to generate a combined data set. This paper mainly focuses on web content mining technique for the generation of covid-19 twitter data set.

Keywords: Web data mining; hyperlinks; usage logs; contents; patterns; web crawling; deep neural networks.

I. INTRODUCTION

The exponential grows of information sources available in the World Wide Web demands for automated tools to find the desired information resource and to track and analyze user behavior using a Web application. This reason causes the necessity of developing server-side and client-side intelligent systems, which can efficient mine for knowledge. “Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services [1]”. In this paper we provide an overview of the different categories, techniques and problems concerning Web mining and we take a closer look at Web content mining for generating the dataset. When collecting server-side data various challenges, like the integration of several data sources such as server access logs, referrer logs or profile information, have to be coped. The process of Web mining can be divided into four subtasks [1]: (i) Finding Resources: the task of retrieving intended Web documents. (ii) Information selection and preprocessing: selection and preprocessing specific information from retrieved Web resources by automated mechanisms. (iii)Generalization: Discovering general patterns at individual Web sites as well as across multiple sites. (iv) Analysis: validation and/or interpretation of the mined patterns.

Now a day’s data over the internet is enormous and increasing frequently day by day. It is must to manage that massive information and display most related queried information on user’s screen. Analyzing and fetching relevant data from large data bases is not possible manually, for this automated extraction tools are required through which user queried data can be fetch from billions of pages over the internet and discovers relevant information. Usually users find data from world wild web WWW by using different search engines like Yahoo, Bing, MSN, Google etc. A lot of user data is also available on the social media applications like Twitter. Data mining is a process of analyzing usable information and extract data from large data warehouses, involving different patterns, intelligent methods, algorithms and tools. This process can help business to analyze data, user behavior and predict future trends. Data mining includes four strategies steps for relevant data extraction. Data source is a set on data in large data base which can have problem definition in it. Data exploration is a step of investigation true information from bulks of unfamiliar data. Third step is modeling, in these different models are designed and then evaluate. At the end tested models are deployed, that occurs in final step of data mining strategies. Organizations can use data mining techniques to change raw data into convenient information. It can also help business to improve their marketing strategies and increase the profit by learning more about customer's behavior.

Web mining is one of the types of techniques use in data mining. The main purpose of web mining is to automatically extract information from the web. For discovering useful data (videos, tables, audio, images etc.) from the web different techniques and tools are used. Information over the internet is huge and increasing with passage to time due to which size of data bases are also growing. Digging knowledgeable information and analyzing the data sets for relevant data is much difficult because data over the internet in not in plain text. It could be unstructured data, multimedia, table, tag.

Twitter is the place where millions go to express their opinions on any topic, and it has also become a place for interactions between celebrities and their fans. Purpose of this paper is to describe web mining, its three different types, tools and techniques. Its main purpose is to generate the data set for Covid- 19 tweets.

II. LITERATURE REVIEW

Data mining is a process of discovering knowledge from data warehouse. This knowledge can be classified in different rules and patterns that can help user/organization to analyze collective data and predicted decision processes [9]. Centralized database of any organization is known as Data warehouse, where all data is stored in a single huge database. Data mining is a method that is used by organization to get useful information from raw data. Software's are implemented to look for needed patterns in huge amount of data (data warehouse) that can help business to learn about their customers, predict behavior and improve marketing strategies.

Web mining is actually an area of data mining related to the information available on internet. It is a concept of extracting informative data available on web pages over the internet [1]. Users use different search engines to fetch their required data from the internet, that informative and user needed data is discovered through mining technique called Web Mining. Different tools and algorithms are used for extraction of data from web pages that includes web documents, images etc. Web mining is rapidly becoming very important due to size of text documents increasing over the internet and finding relevant patterns, knowledge and informative data is very hard and time consuming if it is done manually. Structure (Hyperlinks), Usage (visited pages, data use), content (text document, pages) are included in information gathered through Web mining [2], [5]. Term World Wide Web is related to the combination of web documents, videos, audios etc. Some processes included in web mining are

Information Retrieval is a process of retrieving relevant and useful information over the web. Information retrieval has more focuses on selection of relevant data from large collection of databases and discovering new knowledge from large quantity of data to response user query.IR steps includes searching, filtering and matching [5], [6].

Information extraction is an automatic process of extracting analyzed data (structured). IE is a task that work same like information retrieval but more focuses on extracting relevant facts [5].

Machine Learning is support process that helps in mining data from web. Machine learning can improve the web search by knowing user behavior (interest). Different machine learning methods are used in search engine to provide intelligent web service. It is much more efficient than traditional approach i.e. information retrieval. It is a process that has ability to learn user behavior and enhance the performance on specific task.

III. WEB MINING CATEGORIES

Web Mining is sub categorized into three types as shown in Fig. 1:

- Web Structure Mining

- Web Usage Mining
- Web Content Mining

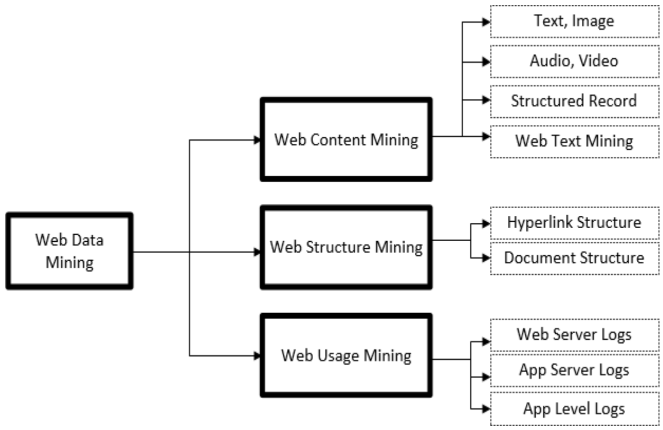

Fig. 1. Web mining taxonomy [8], [9], [15].

Web Mining consists of massive, dynamic, diverse and mostly unstructured data that provides big amount of data. Explosive growth of web leads to some problems like finding relevant data over the internet, observing user behavior. To solve such kind of problem efforts were made to provide relevant data in structure form (table) that is easy to understand and useful for organizations to predict customer's needs [4].

Web Structure Mining

Now a day's massive amount of data is increasing on web. World Wide Web is one of the most loved resources for information retrieval. Web mining techniques are very useful to discover knowledgeable data from web. Structure mining is one of the core techniques of web mining which deals with hyperlinks structure [14]. Structure mining basically shows the structured summary of the website. It identifies relationship between linked web pages of websites. Continues growth of data over the internet become a challenging task to find informative and required data [15]. Web mining is just a data mining which digs data from the web. Different algorithmic techniques are used to discover data from web. Structure mining analyzes hyperlinks of the website to collect informative data and sort out in categories like similarities and relationship. Intra-page is a type of mining that is performed at document level and at hyperlink level mining is known as inter-page mining. Link analysis is an old but very useful method that is way its value increases in the research area of web mining – Structure analysis is also called as Link-mining [16].

Web Structure Mining Algorithms: There are various web structure mining algorithms as mentioned in Table I, the paper describes two of them i.e. Page rank algorithm and HITS algorithms. Both of them focuses on link structure of web and how it gives importance to web pages.

Web Structure Mining Tools: Following mention tools are used for web structure mining. Google PR checker is a tool designed for page rank and is used for Page rank algorithm. It helps to rank of web pages in search engine result. It is simple to find page rank just by pasting website URL and click search – it will show rank of each page of website. Link viewer is used for HITS to visualize analyzes links process.

- Google PR Checker (for PageRank)
- Lin Viewer (for HITS)
- Web Mining Categories Summarization

Web Usage Mining:

Web usage mining also called log mining is a process of recording user access data on the web and collect data in form of logs. After visiting any website user leaves some information behind such as visiting time, IP address, visited pages etc. This information is collected, analyzed and store in logs. Which helps to understand user behavior and later can improves website structure [18]. Web usage mining is a technique that automatically archives access patterns of user and this information is mostly provided by web servers which are later collected in access logs. Logs stores much needed information like URL address, visiting time, Internet Protocol addresses etc. which can help an organization to understand their customer's behavior and insure good service quality. Web usage mining dig and analyze data present in log files which contains user access patterns. Main purpose of web usage mining is to observer user behavior at the time of his interacting with web. There are two types of pattern tracking i.e. general tracking and customized tracking. In general tracking information is collected from web page history. In customized tracking the information is gathered for specific user [19]

Web Usage Mining Algorithms: For usage mining there are numbers of algorithms that can be used as few of them are listed in above Table I. This section will describe three important algorithms i.e. Apriori Algorithm, FP Growth Algorithm and Fuzzy c-means algorithm.

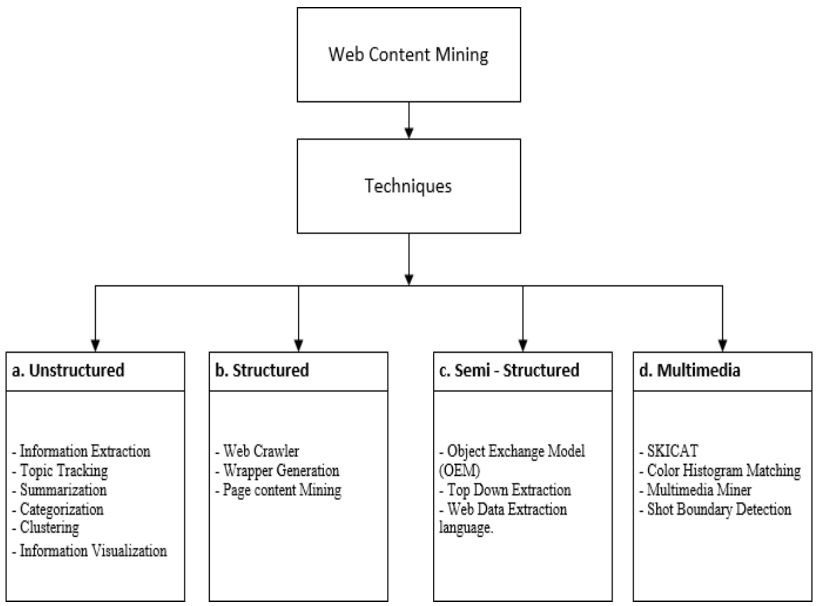

Fig. 2. Web content techniques [7].

Web Usage Mining Tools: Speed Tracer is an analysis tool and use for usage mining. This tool help to discover users surfing behavior and analyze with entries stored in server logs by using different data mining techniques. Cookies are not required for identifying user session, speed tracer used different kind for information like: IP address, URL of page, agent etc. Collection of browsing patterns helps to understand user behavior in a better way [18]. Three types of understanding are generated by speed tracer: User based which refers to user access time duration. Path based relates to process of frequently visited path in web. Group based generates information of repeatedly visited groups of web pages.

Web Content Mining

Content Mining is a process of Web Mining in which needful informative data is extracted from web sites (WWW). Content includes audio, video, text documents, hyperlinks and structured record [1]. Web contents are designed to deliver data to users in the form of text, list, images, videos and tables. Over last few decades the amount of web pages (HTML) increases to billions and still continues to grow. Searching query into billions of web documents is very difficult and time consuming task, content mining extracts queried data by performing different mining techniques and narrow down the search data which become easy to find required user data [3][17].

Web Content Mining Techniques: Web content mining uses different techniques Fig. 2 to dig data. Following are four techniques described used by web content mining.

Mostly in web contents data is in unstructured text form. For extraction of unstructured data, web content mining requires text mining and data mining approaches [5]. Text documents are related to text mining, machine learning and natural language. Main purpose of text mining is to extract previous information from content source [7]. Text mining is a part of web content mining and hence different techniques are used for text data mining from web contents over the internet/website to provide unknown data, some of them are mentioned below:

- Information Extraction
- Summarization
- Information Visualization
- Topic Tracking
- Categorization
- Clustering

Structured is a technique that mines structured data on the web. Structure data mining is an important technique because it represents the host page on the web. Compare to unstructured, in structured data mining it is always easy to extract data [8]. Following are some techniques used for structured data mining:

- Web Crawler
- Page Content Mining: Scraping
- Wrapper Generation

Semi-Structured is form of structured data but not full, text in semi-structured data is grammatical. Its structure is Hierarchical, not predefined. Representation of semi structured data is in form of tags (such as HTML, XML). HTML is an intra-document structure case [4]. Techniques used to extract semi-structured data are:

- OEM - Object Exchange Model
- Web Data Extraction Language
- Top Down Extraction [5]

Traditionally computing data was considered as text and numbers but now a days there are different computing data types of multimedia data like videos, images, audios etc. This mining process is used for extracting interesting multimedia data sets and also converted data set types in to digital media [11]. Techniques uses for multimedia data mining are:

- Multimedia Miner
- Shot Boundary Detection
- SKICAT
- Color Histogram Matching [10]

TABLE I. SUMMARIZATION TABLE FOR WEB DATA MINING CATEGORIES

Web Mining Categories	Techniques	Tools	Algorithms
Web Content Mining	- Unstructured Data Mining - Structured Data Mining - Semi – Structure Data Mining - Multimedia Data Mining	- Screen Scaper - Mozenda - Automation Anywhere7 - Web Content Extractor - Web Info Extractor - Rapid Miner	- Decision Tree - Naive Bayes - Support Vector Machine - Neural Network
Web Structure Mining	- Link-based Classification - Link-based Cluster Analysis - Link Type - Link Strength - Link Cardinality	- Google PR Checker - Link Viewer	- Page Rank Algorithm - HITS algorithms (Hyperlink Induced Topic Search) - Weighted Page Rank Algorithm - Distance Rank Algorithm - Weighted Page Content Rank Algorithm - Webpage Ranking Using LinkAttributes - Eigen Rumor Algorithm - Time Rank Algorithm -Tag Rank Algorithm - Query Dependent Ranking Algorithm
Web Usage Mining	- **Data Preprocessing** • Data Cleaning • User & Session Identification - **Pattern Discovery** • Statistical Analysis • Association Rules • Clustering • Classification • Sequential Patterns - **Pattern Analysis** • Knowledge Query Mechanism • OLAP (Online Analytical processing) • Intelligent Agents	- **Data Preprocessing Tools** • Data Preparator • Sumatra TT • Lisp Miner • SpeedTracer - **Pattern Discovery Tools** • SEWEBAR-CMS • i-Miner • Argunaut • MiDas(Mining In-ternet Data for As-sociative Sequenc-es) - **Pattern Analysis Tools** • Webalizer • Naviz • WebViz • WebMiner • Stratdyn	- **Association Rules** • Apriori Algorithm • Maxi-mal Forward References • Markov Chains • FP Growth • Prefix Span - **Clustering** • Self-Organized Maps • Graph Partitioning • Ant Based Technique • K-means with Genetic algorithms • Fuzzy c-mean Algorithm - **Classification** • Decision Trees • Naïve Bayesian Classifiers • K-nearest Neighbor Classifiers • Support Vector Machine - **Sequential Patterns** • MIDAS (Mining Internet Data for Association Sequences) algorithm

1) Web Content Mining Algorithms: Multiple techniques are used by web mining to extract information from huge amount of data bases. There are different types of algorithms that are used to fetch knowledge information, below are some classification algorithms are described:

Decision tress is a classification and structured based approach which consist of root node, branches and leaf nodes. It is hierarchical process in which root node is split into sub branches and leaf node contains class label. Decision tress is a powerful technique [10].

Naive Bayes is an easy, simple, powerful algorithm for classification and also known as Native Bayes classifier. It is based on Bayes' Theorem. From predefined dataset values, probabilities are calculated for each class by counting combinations on values. Most likely class is the one with highest probability [12].

Support Vector Machine is a well-known and simple machine learning and classification algorithm. SVM is a method that can be used for linear and non-linear data sets [10]. Optimal separating hyper plane (decision boundary) is just a line that is used to draw to separate the two classes depends on the different classification features.

Neural network is another web content mining approach which use back propagation algorithm. The algorithm consist of multiple layers i.e. input layer, some hidden layers and then output layer, each feeds the next layer till last layer (output). Neuron is the basic unit of neural network. Inputs are fed simultaneously to units. From input layer, inputs are simultaneously feeding to hidden layers. Usually there is one hidden layer but numbers of hidden layers are arbitrary [10]. Last hidden layer fed the input and make up the output layer.

IV. Research Methodology with Neural Networks

The micro-blogging platform Twitter has become one of the best-known social networks on the internet (Kowald,2017). Twitter is considered among the first social networking sites to use the hashtag concept. In twitter, users are liberally permitted to allocate hashtags to their tweets, which are nothing but a string whose prefix is the hash symbol, and are used to catalog their posts and highlight certain content (Kabakus,2019). The hashtag is considered a community-driven principle that adds extra context to tweets. This method assists in searching, rapidly propagates the topic from millions of users, and allows users to link into specific discussions (Li, Y.;2019). Below figure 3 shows the architecture of the combined approach of web content mining and neural networks. We name it as Semantic Enhanced Deep Neural Network Framework (SEDNNF)

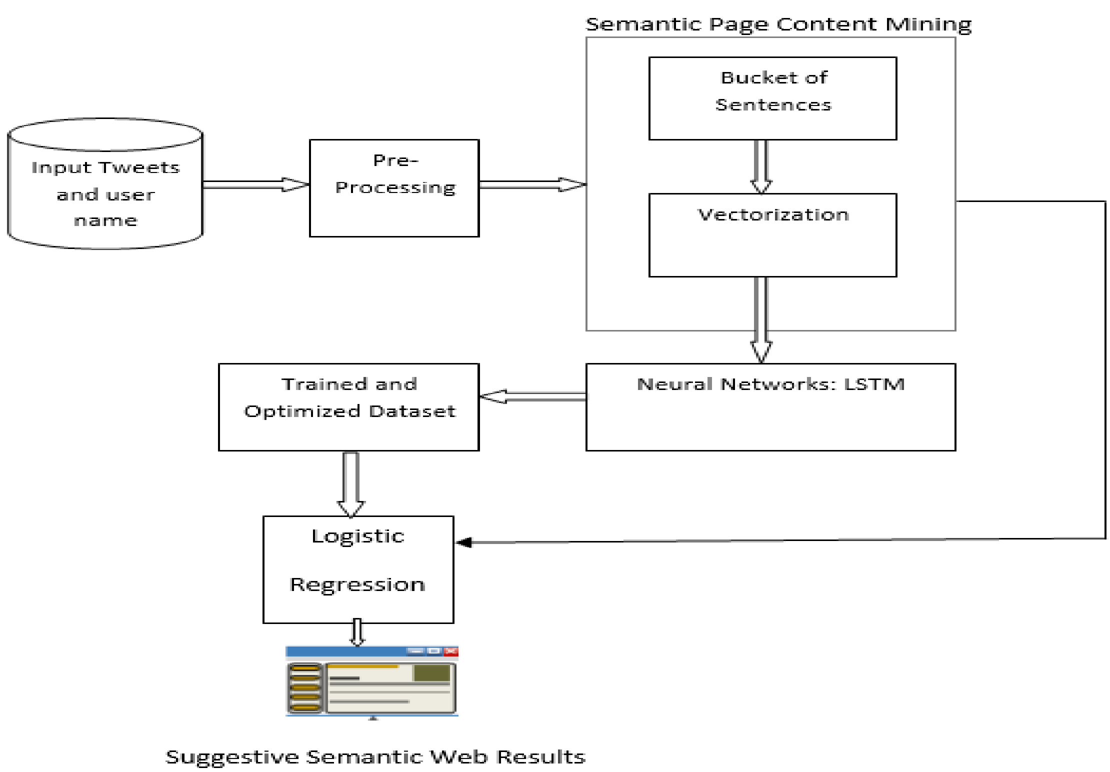

Figure 3. Architecture of Semantic Enhanced Deep Neural Network Framework (SEDNNF)

In the proposed methodology, we are using the neural network algorithm LSTM (Long Short-Term Memory) for the latest tweet generation, based on the hashtags used in the tweets that follow Covid-19 specific tweets from the user. To make the tweet searching process easy, twitter users use hashtags for classifying their tweets. In this paper, the process of hashtag recommendation is carried out using a Recurrent Neural Network, in such a way that the training process of deep learning classifier is carried out using the LSTM algorithm (Ben-Lhachemi,2019).

LSTM (Long Short-Term Memory) is a type of Recurrent Neural Network (RNN). We chose to create an LSTM model to generate new tweets for the set of users because this particular model learns and picks up on style, language and grammar because an LSTM is able to detect and learn long-term dependencies. In learning the sequences of the input text, the model is then able to generate new tweets.

LSTMs use a process of learning that can be described as a looping effect, meaning that the network interprets an input and generates an output as shown in figure 4. This output and its information is then passed to the next step within the network. This is essential because it allows the model to pick up on patterns within the text in order to generate better outputs similar to the style of text in question.

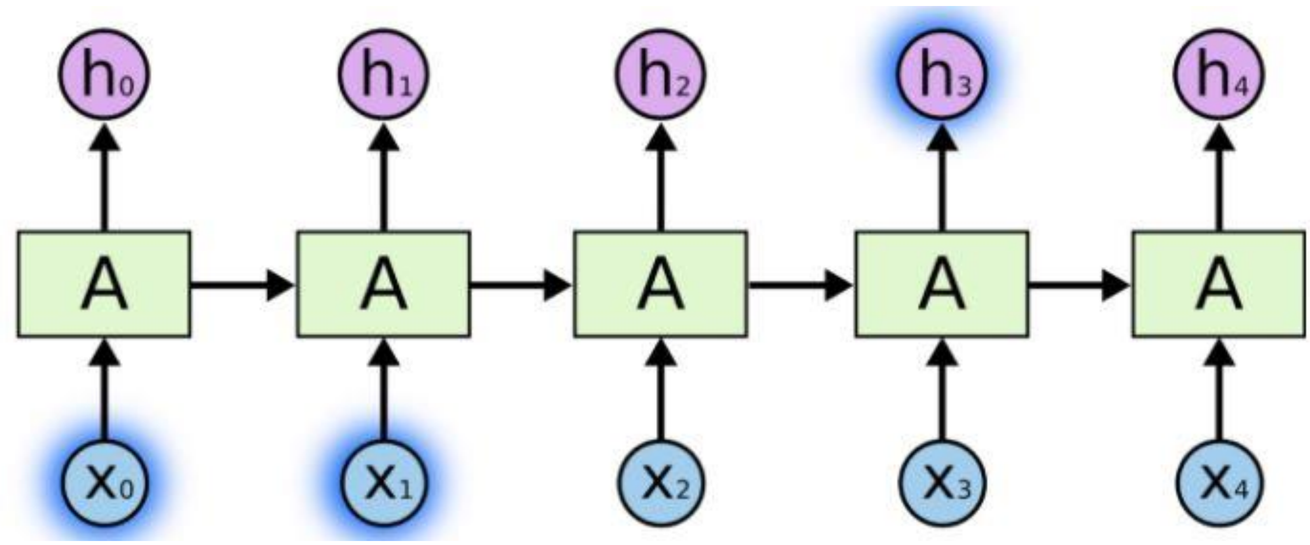

Figure 4: Recurrent Neural networks

In order to generate latest covid-19 tweets using an LSTM network, the data set of older covid-19 tweets in 2020 are taken from Kaggle. The data from web is scrapped and updated in the new data set till the month of September 2021 that is trained using the older data set. As shown in figure 5, the data set has 1, 70,000 entries with 13 fields namely: “user name, user location, user description, user created, user followers, user friends, user favorite, user verified, date, text, hash tags, source, is re-tweeted”. Tweets and user names from the CSV were transferred into a pandas file. Since there are disparities within the tweets, data cleaning becomes important but without lowering the amount of data that would be used in training the model. Only removal of stop words in the tweets was done. A stop word is the word that appears often and does not contain any meaning. Stop word removal is the process of removing these words, including ‘the’, ‘a’, ‘is’, ‘an’, and so on. As the stop word does not have any sentiment, it is very significant to remove these words from the tweet. The stop word removal process increases the performance of hashtag recommendation by reducing the dimensionality rate.

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‰áŽ¥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 č	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–Šï¸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	ÐÐ¾Ð²Ð¾Ñ€Ð	ðŸŽ¼ #ÐÐ¾Ð²Ð¾	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 i	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	Creativegms	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»@Patty	['COVID19	Twitter W	FALSE
15	Africa Youth	Africa	Official account c	13-05-2019 06:27	830	254	3692	FALSE	25-07-2020 12:26	Let's all	['COVID19	Twitter W	FALSE
16	DailyaddaaN	New Delhi	Breaking news al	22-10-2016 09:18	546	29	88	FALSE	25-07-2020 12:26	Rajasthan Government 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	ChennaiCityNow		Individual tweet	26-04-2009 09:38	3987	53	749	FALSE	25-07-2020 12:26	July 25	['COVID19	Twitter fo	FALSE
19	marc goovae	Brussels	Progressive minc	13-06-2009 13:48	283	1432	1546	FALSE	25-07-2020 12:26	Second wave c	['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 I	#AFRICA #MEI	Latest #Africa & #	22-02-2011 09:09	10661	6	2037	TRUE	25-07-2020 12:26	Coronavirus - South Afric		Africa Nev	FALSE
24	Micah Pollak	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 5: Covid-19 twitter data set of year 2020

Next step is semantic page content mining where tweets were divided into sequences of equal lengths and placed into a list called sentences. These sentences will be used as training examples for the model (Khabiri,2012).

Vectorization is the next step. Vectorization is the process of converting an algorithm from operating on a single value at a time to operating on a set of values (vector) at one time. Modern CPUs provide direct support for vector operations where a single instruction is applied to multiple data (SIMD).

Here, generate 3 lists of words namely: vocabulary list, thematic list and dictionary list.

Vocabulary list: set of all words/ characters from the bucket of sentences.

Dictionary list: set of semantic and recurring words after performing vectorization.

Thematic list: set of hash tagged words from the dictionary list.

We now generate the 2D matrix. For example, consider a case where the tweets are taken as input and divided into sentences.

S1: The #Covid-19 virus has spread #worldwide to a great extent.

S2: The spread of #Covid-19 has got the world into great #loss.

The vocabulary list will be created as follows:

Voc: {#Covid-19 virus spread #worldwide great extent world #loss}

Vector for S1 is created by taking value 1 if the word from S1 is present in the vocabulary and 0 if the word is absent in the vocabulary.

V1:

1	#Covid-19
1	virus
1	spread
1	#worldwide
1	great
1	extent
0	world
0	#loss

V2:

1	#Covid-19
0	virus
1	spread
0	#worldwide
1	great
0	extent
1	world
1	#loss

We merge these common words from the vector and form the Dictionary

D{#Covid-19 spread great }

The thematic list is finally prepared by extracting the hashtag words from the dictionary.

T{#Covid-19}

Below is the figure 6 explaining the above steps:

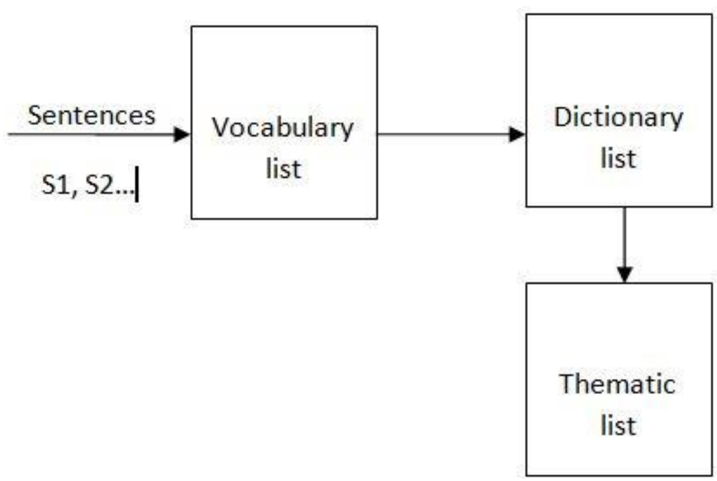

Figure 6: Vocabulary list, Dictionary List, Thematic List

Now that the data has been scraped and prepared for training, it’s time to build the LSTM model.

Deep LSTM (Gers, 2000) is employed to determine the tweets with hashing content, considering the keyword set features as input. A deep neural Network constructed by using a number of layers within the LSTM framework is called a Deep LSTM. It takes advantages of the benefits of both deep neural Networks and LSTM Networks to obtain the hashing score. It consists of internal state cells, which act as memory cells. The output of the Deep LSTM is the hashing score, which is computed based on the cell states. The working process of this Network solely depends on the memory cells.

The input node receives the input feature f , where f = { f1, f2, f3, f4, f5, f6, f7, f8}, at the input layer.

$$Q_F = \tanh(f * K_{OF} + M_{F-1}K_{QM} + Z_Q), \qquad (1)$$

where tanh denotes the hyperbolic tangent function, QF denotes the input node at time F, f denotes the input feature, KOF indicates the weights among the input layer and input nodes of the memory cell, MF−1 represents the input of the hidden state, KQM denotes weight matrix for the hidden state, and ZQ indicates the bias of input node.

The input gate, YF, is equivalent to that of the input node, and it receives the input feature f . The input gate blocks information flow between the nodes. The input gate is represented as

$$YF = X(f * KOF + MF-1KQM + ZY), \qquad (2)$$

where YF denotes the input gate, ZY denotes the bias of the input gate, and X specifies the sigmoid activation function.

The internal state, λF, is a node with a self-loop recurrent edge consisting of a linear activation function and a unit weight, such that the internal state can be represented as

$$\lambda F = YF \,\Theta\, QF + \lambda F-1, \qquad (3)$$

where λF denotes the internal state at time F. The forget gate is the sub-unit used for re-initiating the internal state of the memory cell, which is computed as

$$\beta F = X(f * KM\, f + MF-1K\beta M + Z\beta), \qquad (4)$$

where βF denotes the forget gate, KβM denotes weights among the forget gate and the hidden states, and KM f represents the weights among the forget gate and the input layer.

Finally, the output gate, γF, is expressed as

$$\gamma F = X(f * K\gamma\, f + MF-1K\gamma M + Z\gamma), \qquad (5)$$

where Kγf denotes the weights among the output gate and the input layer, KγM represents the weights among the output gate and the hidden states, and Zγ denotes the bias of the output gate. The output of memory cell is calculated as

$$MF = \tanh(\lambda F)\, \Theta\, \gamma F\ , \qquad (6)$$

where $\lambda F = QF\ \Theta\ f + \lambda F-1\ \Theta\ \beta F$ and Θ denotes the XNOR operator. The score computed using the Deep LSTM is termed the hashing score, and is represented as f9 with the dimension of [V × 1]. It is used to perform the hashtag recommendation, by passing it as an input to the Deep Residual Network, in addition to the features of the keyword set. Figure 7 shows the structure of the Deep LSTM classifier.

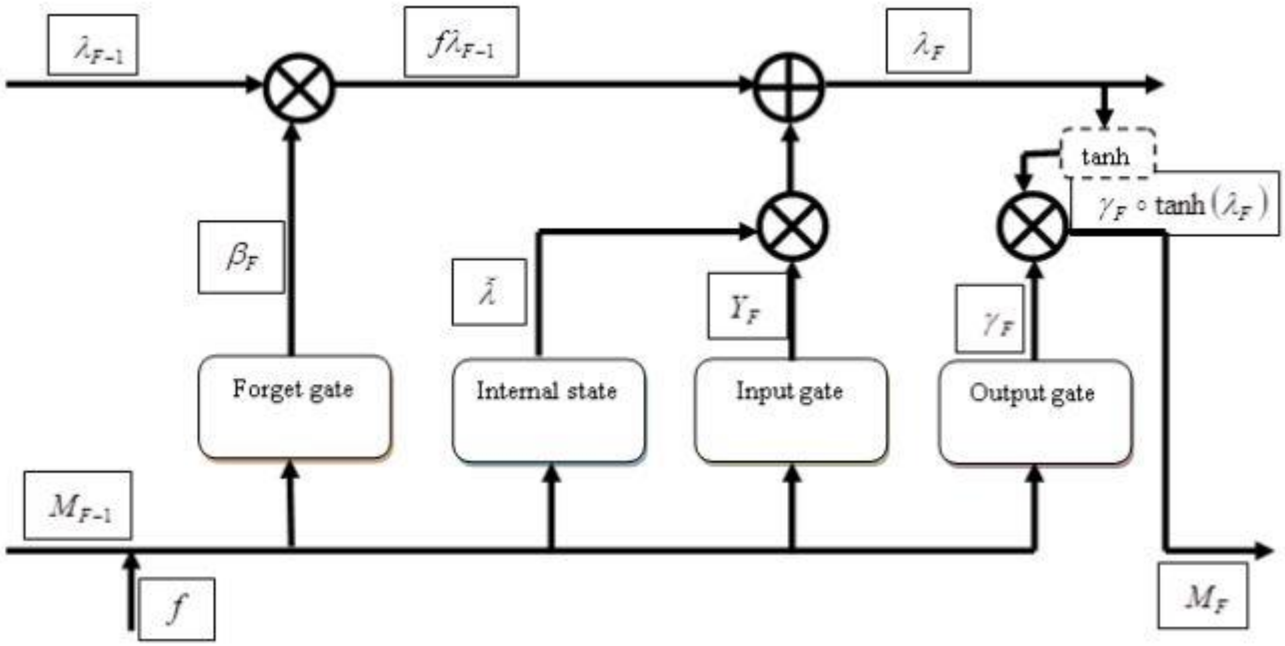

Figure 7: Structure of Deep LSTM classifier.

V. Logistic regression for mining the sentiment

After training the model using LSTM the input is given to logistic regression algorithm which also takes the previously prepared semantic data, so as to perform the sentiment analysis for getting suggestive semantic web results.

Sentiment analysis is a technique that allows us to classify if a text, tweet or comment is either positive, neutral or negative. Today's technology enables users to express their emotions and thoughts more openly on social platforms than ever before. So, Sentiment Analysis has become a mandatory tool for every business to understand user sentiment and gauge their performance to tailor their products and services according to user needs, thus making systems more efficient.

Logistic regression is a supervised machine learning technique for classification problems. Supervised machine learning algorithms train on a labeled dataset along with an answer key which it uses to train and evaluate its accuracy. The goal of the model is to learn and approximate a mapping function f(Xi) = Y from input variables {x1, x2, xn} to output variable(Y). It is called supervised because the model predictions are iteratively evaluated and corrected against the output values, until an acceptable performance is achieved. As a part of building sentiment classifier using logistic regression, we train the model on twitter sample dataset.

Initially, we classify whole data set into two types (Training and Testing sets). The former one contains 3 columns namely id, label and tweets. Label column contains binary values such as 0 and 1. We proceed by allocating labels to our training set where label 1 denotes a negative tweet whereas label 0 denotes not a negative tweet. Now, we build a model using a logistic regression algorithm. After training this model, it predicts the negative and non-negative tweets in the testing dataset by labeling them with 0 or 1. In our case, we used the f1 score as an evaluation criterion.

Sentiment classification involves following stages:

1. Text pre-processing and cleaning

This is considered as important phase because it prepares the unstructured data for processing. If we have not performed this step properly, then there is a high probability that we may deal with scattered and inaccurate information. So, the goal of performing this phase is to remove unrelated text from the twitter posts like special characters and punctuation which does not add any value to calculate sentiment.

(i) Removing Twitter Handles: When collecting the tweets all the twitter accounts are concealed because of privacy issues. So, we wipe all the twitter account usernames(@user) out of all the gathered tweets since these are not necessary for finding the sentiment.

(ii) Removing punctuations, numbers and special characters: In this step, excluding hashtags and characters we replace every other thing with spaces.

(iii) Removing Short words: In this step, we should take certain measures in choosing the word length meticulously to remove. So, we choose the word length having length of 3 or less to remove from the tweet. For example, words like “hmm”, “oh” which do not have any sentiment.

(iv) Tokenization: Tokens are defined as individual words and the method of breaking a string into tokens is called tokenization. For example, let us take a tweet “He played good cricketing shots in that match”. After Tokenization it looks like [‘He’, ‘played’, ‘good’, ‘cricketing’, ‘shot’, ‘in’, ‘that’, ‘match’]

(v) Stemming: The process of removing suffixes from a word is performed in the stemming phase. For example, sing, singer, singing, sings all come under the root word sing.

2. Visualization from tweets

For this purpose, we use word cloud as a means for representing the key words from the above classified tweets. In this technique, maximum number of used words are displayed in bigger size and least occurring one in the tiny size.

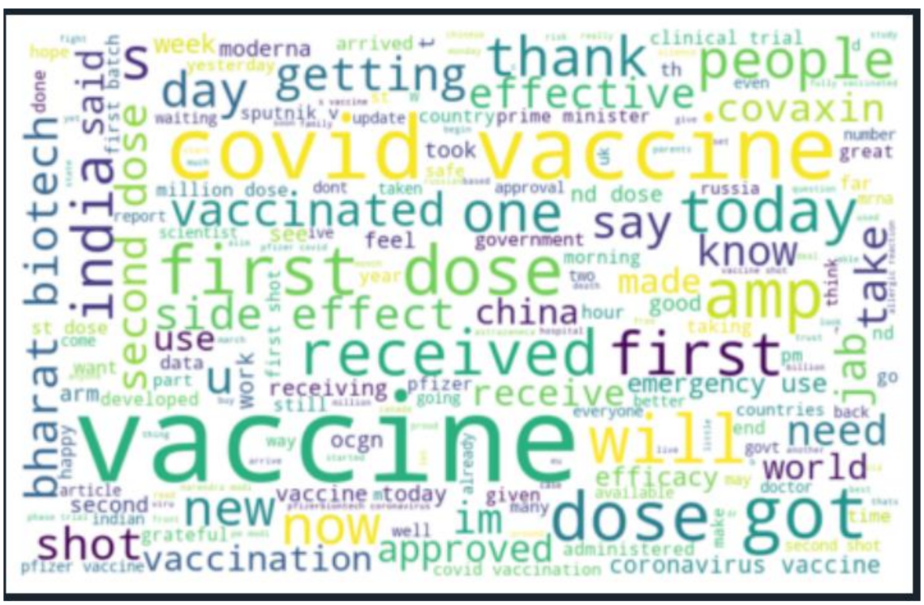

Figure 8: Word Cloud Representation

3. Feature extraction technique

In order to examine pre-processed information, these must be transformed into further stage. There are many ways available for extracting features. They are

(1) Bag of words Approach: This technique is used to show the content into mathematical format. Let us assume a corpus K has R documents {r1,r2,r3...rN} and M distinctive tokens taken from the corpus K. The M tokens generates a matrix in R $\times$ M format. The rows in matrix hold the occurrence of tokens in R(i).

(2) TF IDF Method: TF is denoted by term frequency whereas IDF is denoted by inverse document frequency. This method is also dependent on frequency but can be distinguished from the former method, in such a way that it sees for occurrence of a word in the whole corpus. This approach assigns lowest weight to the most repeated words in the corpus and give preference to words which have occurred very less times in corpus. The significant phrases used in TF-IDF are:

• TF = (the count of the term t appears in a text document)/(Number of terms in the document)

• IDF = log(D/d), where, D is the number of documents and the number of documents a term t has appeared is referred to as d.

• TF IDF = TF*IDF

VI. Outputs, Evaluation and Comparison

The below outputs are generated from the above proposed model.

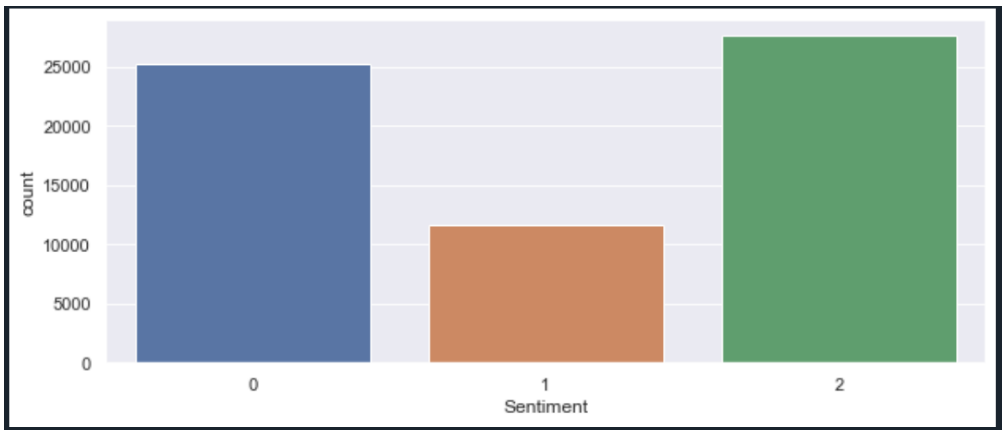

Figure 9: Sentiment analysis on positive, neutral & negative polarity.

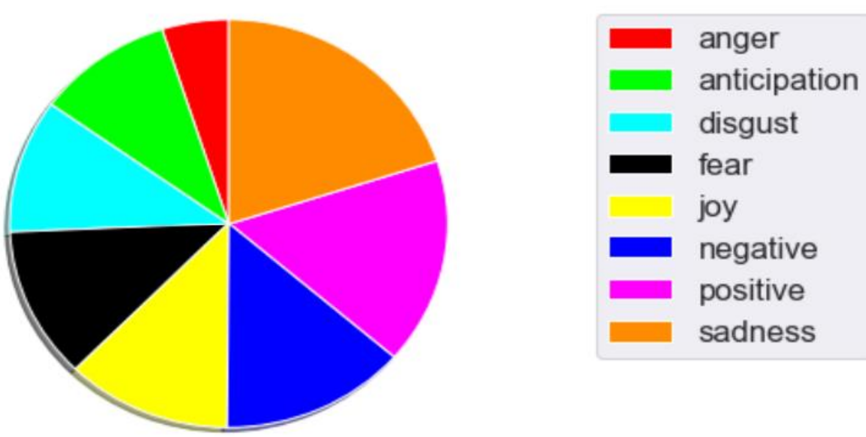

Figure 10: pie chart on different sentiments of tweets.

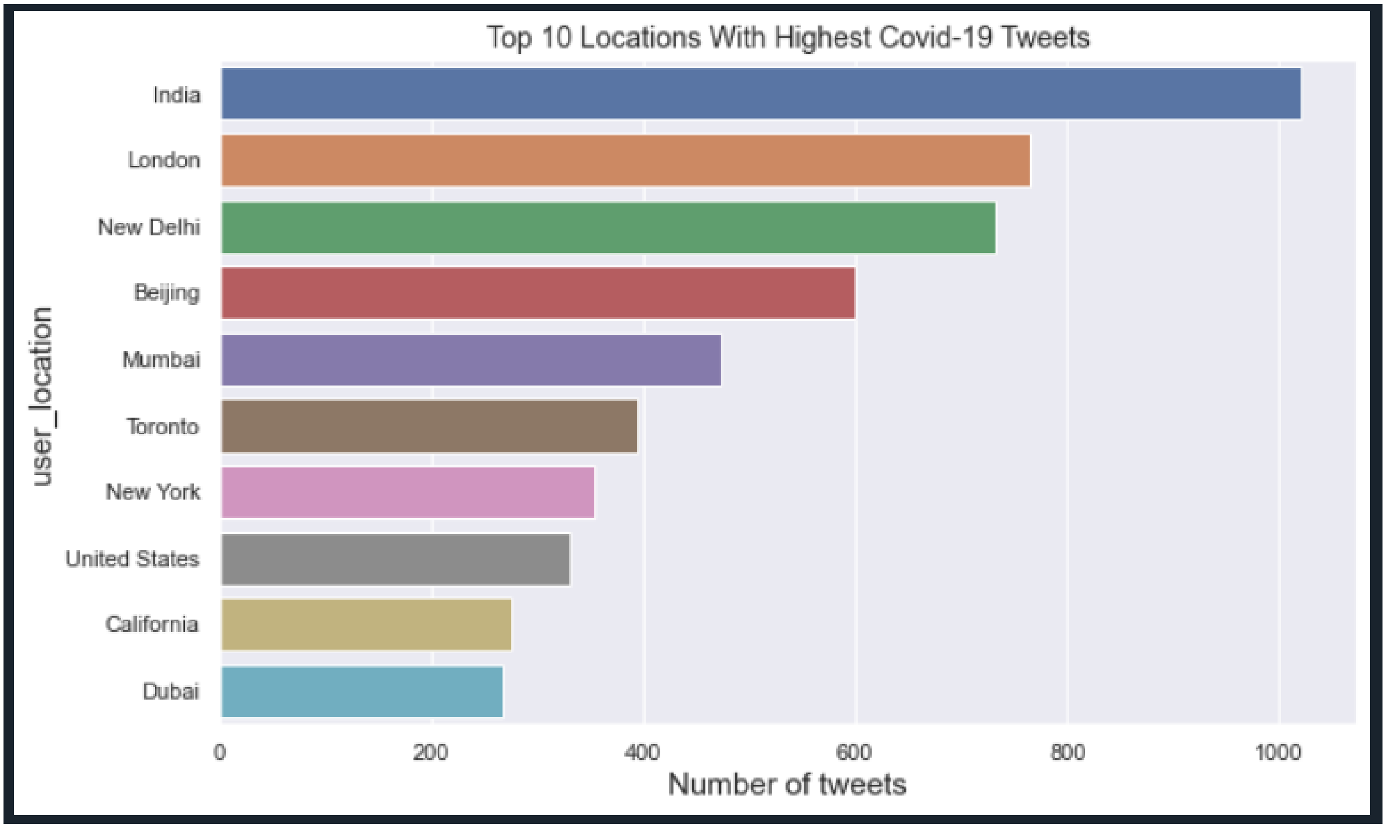

Figure 11: Top 10 locations with highest Covid-19 tweets

For evaluating the performance of the considered architecture approaches we used the following metrics: accuracy, precision, recall, and true negative rate (TNR).

For defining these metrics, we need to first define the following particular terms. The number of true positives (TP) is the number of samples which are labeled as positive by our model and are also annotated as positive by the creators of the dataset. The number of true negatives (TN) is the number of samples labeled as negative and annotated as negative. The number of false positives (FP) is the number of samples labeled as positive, but annotated as negative. Finally, the number of false negatives (FN) is the number of samples which are labeled as negative, but are annotated as positive. In the case of our binary classification task, TP and FP refer to the positive class, while TN and FN refer to the negative class. However, in the general N-class scenario, these terms can refer to any of the classes.

Equations (1) define the metrics. The ratios are multiplied by 100 to present the results in percentages. The accuracy (Acc) is the percentage of samples which were correctly labeled by our model out of the total number of evaluated samples.

$$Acc = ((TP + TN) / (TP + TN + FP + FN)) * 100 \quad (1)$$

Table 2: Comparison between performance of SEDNNF

Method	Training data set	Accuracy (%)
SVM	Training on the same covid-19 data set	86
SEDNNF	Training on the same covid-19 data set	92

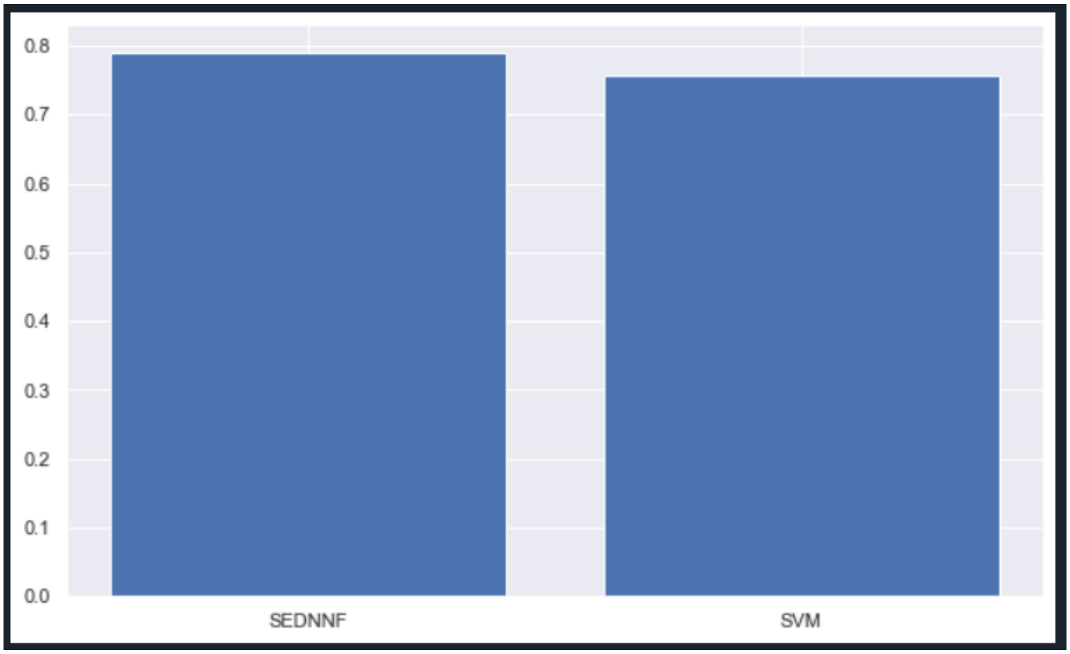

Figure 12: Comparison between performance of SEDNNF

VII. Conclusion and Future Scope

The increase in popularity of social networks created an environment for micro blogging in which people can express ideas and opinions on any topic and their posts can reach a very large number of followers. Twitter is one of the most popular social networks and a very good source of data. This research work has found a combined approach of web content mining and neural networks for Covid -19 twitter data creation and mined the opinion for predicting the impact of covid-19 on society in terms of socio-political and economic status. As it is observed from figure 9, the public sentiment tends to be more negative than positive. Specifically, in figure 10, a pie chart is prepared on the basis of emotions in terms of joy, sadness, fear, anticipation, negative, disgust, anger, and the rate of negative emotions is higher than positive. Thus, we conclude that there was a negative impact on socio-political and economic status.

From fig 11, it is also analyzed that most of the tweets related to covid-19 were done from India, specifically, Mumbai and Delhi, which ultimately have negative sentiments. From table 2 and fig 12, we also observed that there is a significant difference in the results when using SVM and SEDNNF in terms of accuracy, SEDNNF has better accuracy.

The future scope of the work can be data security on web platforms including cloud.

REFERENCES

[1] Anurag Kumar and Ravi Kumar Singh, "Web Mining Overview, Techniques, Tools and Applications: A Survey," International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 12, pp. 1543-1547, December 2016.
[2] Simranjeet Kaur and Kiranbir Kaur, "Web Mining and Data Mining: A Comparative Approach," International Journal of Novel Research in Computer Science and Software Engineering, vol. 2, no. 1, pp. 36-42, January - April 2015.
[3] Ahmad Tasnim Siddiqui and Sultan Aljahdali, "Web Mining Techniques in E-Commerce Applications," International Journal of Computer Applications, vol. 69– No.8, pp. 39-43, May 2013.
[4] Kshitija Pol, Nita Patil, Shreya Patankar, and Chhaya Das, "A Survey on Web Content Mining and extraction of Structured and Semistructured data," Emerging Trends in Engineering and Technology, pp. 543-546, July 2008.
[5] R. Malarvizhi and K Saraswathi, "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study," International Journal of Computer Trends and Technology (IJCTT), vol. 4, no. 8, pp. 2940- 2945, Augest 2013.
[6] Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey," SIGKDD Explorations, vol. 2, no. 1, pp. 1-15, July 2000.

[7] Faustina Johnson and Kumar Santosh Gupta, "Web Content Mining Techniques: A Survey," International Journal of Computer Applications (0975 – 888), vol. Volume 47– No.11, pp. 44-50, June 2012.
[8] Abdelhakim Herrouz, Chabane Khentout, and Mahieddine Djoudi, "Overview of Web Content Mining Tools," The International Journal of Engineering And Science (IJES), vol. 2, no. 6, June 2013.
[9] Claus Pahl and Dave Donnellan, "Data Mining Technology for the Evaluation of Web-based Teaching and Learning Systems," 7th Int. Conference on E-Learning in Business, Government and Higher Education, October 2002.
[10] Anurag kumar and Kumar Ravi Singh, "A Study on Web Content Mining," International Journal Of Engineering And Computer Science, vol. 6, no. 1, pp. 20003-20006, January 2017.
[11] Dr. S. Vijayarani and Ms. A. Sakila, "MULTIMEDIA MINING RESEARCH – AN OVERVIEW," International Journal of Computer Graphics & Animation (IJCGA), vol. 5, pp. 69-77, January 2015.
[12] Tina R. Patil and Mrs. S. S. Sherekar, "16. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," International Journal Of Computer Science And Applications, vol. 6, pp. 256-261, April 2013.
[13] Kowald, D.; Pujari, S.C.; Lex, E. Temporal effects on hashtag reuse in twitter: A cognitive-inspired hashtag recommendation approach. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 1401–1410.
[14] Miguel Gomes da Costa Júnior and Zhiguo Gong, "Web Structure Mining: An Introduction," International Conference on Information Acquisition, pp. 590-595, June 27 - July 3 2005.
[15] Anurag Kumar and Kumar Ravi Singh, "A Study on Web Structure Mining," International Research Journal of Engineering and Technology (IRJET), vol. 04, no. 1, pp. 715-720, January 2017.
[16] B. L. Shivakumar and T. Mylsami, "SURVEY ON WEB STRUCTURE MINING," ARPN Journal of Engineering and Applied Sciences, vol. 9, pp. 1914-1923, October 2014.
[17] Monica Sehgal, "Analysis of Link Algorithms for Web Mining," International Journal of Scientific and Research Publications, vol. 4, no. 5, May 2014.
[18] Pranit Bari and P.M. Chawan, "Web Usage Mining," Journal of Engineering, Computers & Applied Sciences (JEC&AS), vol. 2, pp. 34- 38, June 2013.
[19] Kamika Chaudhary and Santosh Kumar Gupta, "Web Usage Mining Tools & Techniques: A Survey," International Journal of Scientific & Engineering Research, vol. 4, no. 6, pp. 1762-1768, June 2013.
[20] Kabakus, A.T.; Kara, R. “TwitterSpamDetector”: A Spam Detection Framework for Twitter. Int. J. Knowl. Syst. Sci. (IJKSS) 2019, 10, 1–14.
[21] Ben-Lhachemi, N.; Boumhidi, J. Hashtag Recommender System Based on LSTM Neural Reccurent Network. In Proceedings of the 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, Morocco, 28–30 October 2019; pp. 1–6.
[22] Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. Neural Comput. 2000, 12, 2451–2471.
[23] Khabiri, E.; Caverlee, J.; Kamath, K.Y. Predicting semantic annotations on the real-time web. In Proceedings of the 23rd ACM Conference on Hypertext and Social Media, Milwaukee, WI, USA, 25–28 June 2012; pp. 219–228.
[24] Li, Y.; Liu, T.; Hu, J.; Jiang, J. Topical co-attention networks for hashtag recommendation on microblogs. Neurocomputing 2019, 331, 356–365.
[25] Belhadi, A.; Djenouri, Y.; Lin, J.C.W.; Cano, A. A data-driven approach for Twitter hashtag recommendation. IEEE Access 2020, 8, 79182–79191.