EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Design of Novel ETL Model to Analyse Corona Virus Data

Amit Kumar Dewangan[1,*], S.M. Ghosh[2], Akhilesh Kumar Shrivas[3]

[1]Department of Information Technology, Guru Ghasidas Vishwavidyalaya, Bilaspur. India.
[2]Department of Computer Science and Engineering, Dr. C.V. Raman University, Kota, Bilaspur. India.
[3]Department of Computer Science and Information Technology, Guru Ghasidas Vishwavidyalaya, Bilaspur. India.

Abstract

INTRODUCTION:
The corona disease was first recognized in 2019 in Wuhan, which is the capital of China's Hubei-province, and from then it continued spreading and as a result declared as a pandemic by all nations. The COVID-19 virus has different effects on people in various ways. It is a kind of respiratory disease. The confirmed cases are increasing day to day in India, which leads to complete lockdown throughout the nation.
OBJECTIVE:
The objective of this research is to design a novel Extract-Trandform and Load NETL model to analyse covid-19 data in india.
METHODS:
The extraction of useful information from a large database is a well-connected research field of text mining. This paper is proposed a novel extract-transform-load ETL model to process the COVID-19 data of India to get the exact recovery data from the multiple data sources from different states of India. In this, a knowledge-based model that generate knowledge based on three different module split, validation, and join is discussed.
RESULTS:
The outcomes of the proposed NETL process are, output file which has the description of total positive cases, active cases, recovery cases, and death rate, based on different regions. The analysis of NETL is done based on accuracy, failure count, and execution time. The proposed NETL process is more accurate and taking less compilation time with minimum failure count as compared with existing models.
CONCLUSION:
To analyze the coronavirus data in India, a novel ETL (NETL) model is proposed. In this model, a total of 9 CSV files is processed as input files to get different results in different categories. This model is having three modules namely splitting, verification, and join. The dataset is split into based on its coupling attributes and then joined with a single value to produce the updated results as per the current dataset. The last stage of this process is to join the data which is generated through splitting. The proposed NETL model is more accurate as compared with existing ETM models.

Received on 20 May 2020; accepted on 14 July 2020; published on 22 July 2020

Keywords: Corona Virus, Text Mining, Data Analytics, ETL, Covid-19, Pandemic.

doi:10.4108/eai.13-7-2018.165671

1. Introduction

Self-information extraction is consistently been a significant application and research area since the origin of digital records. Therefore, classification and clustering of text is a need because of the extremely huge measure of content archives that we need to manage in day to day life. All in all, content order incorporates the text characterization based on topic, keyword, and cluster which have common properties. Moreover, text mining is a technique in which a document is classified under some predefined

*Corresponding author. Email: amit.nitrr@gmail.com

properties. In a simple mathematical way, it can be represent like, if t_i is a text document and has the set of text document $T_d = l_1, l_2, l_3,.....l_n$, then a classification of text document will be assign l_j to a text document T_d.

Texts may in the form of any news, scientific reports, healthcare information, reviews of any product, social media timeline information, etc. The text is extracted through text mining, where machine learning helps to test and learn the existing text to do the classifications. In continues to this, there is a need for a dataset for supervised machine learning, which may lead to assigning the text document more than one classification known as ranking classification. In the paper, the COVID-19 healthcare dataset has been taken to the consideration of text mining to test and learn the novel approach of text mining with a machine learning technique. The different text mining area is discussed below:

1.1. Information Retrievel

It is nothing but a document retrieval. It helps to narrow point the set of text documents from the same problem. It applies a very complex algorithm.

1.2. Data Mining

It works on a specific pattern in a text document. It may find hidden information for the set of text documents. These tools can analyze and predict text behavior and produce knowledge-based text information for decision making.

1.3. Natural-Language-Processing NLP

This old process is focused on the challenging problem that can study speaking languages. The machine/computer can easily understand the human language through NLP. The leading role of NLP in the mining of text is to produce input from speaking language.

1.4. Information Extraction

Information Extraction is a process that automatically extracts the structured text from unstructured text documents. In this, mostly the human languages are extracted language text by using NLP.

1.5. Text-Mining Process

This is a process in which a batch of activites acts to perform the mining of information, which is presented in Figure 1.

The above-mentioned Figure 1. involves six phases to mine the information for the text document. In this research, the design of a novel ETL model (NETL) is proposed, which is more focused on the transformation process of information as compared with the research done to date. For training and testing, we have taken dataset from Indian Government website which is collecting from various states by World-Health-Organization and produce authentic data https://api.covid19india.org/csv/. This dataset is openly available on the mentioned Indian Government website and provide authentic information.

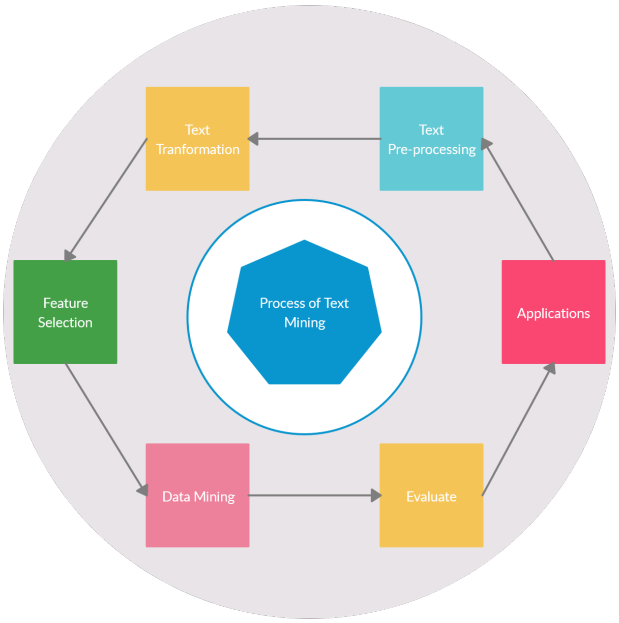

Figure 1. Process of Text Mining

2. Literature Review

The literature of this research is divided into two phases, in the first phase, the different text-mining process is studied and analyzed, and in the second phase, the COVID-19 pandemic is studied based on the current situation in India and analyzed based on different situation.

2.1. Study on ETL Techniques

ETL used for visual analysis, which drives you to provide you the shape of information. Author Costello [1] discussed Tablue tool in his research, where author Galici [2] used the existing ETL model to collect the information from different sources in the blockchain. Author Mallek [3] used the existing ETL model in Big Data to analyze Twitter and Facebook data. Author Awiti [4] proposed a novel ETL model based on relational algebra and Business-Process-Modeling-Notation BPMN at the conceptual level, while author Semlali [5] presented SAT-ETL for satellite big-data

Table 1. Comparative Analysis of recent ETL models based on Data analysis, preperation and report functions.

Functions	[1]	[2]	[3]	[4]	[5]	NETL
Triger entry	-	✓	-	✓	✓	-
Enrichment rule	-	-	✓	✓	-	✓
Selection criteria	-	✓	-	-	✓	✓
Coupling rule	-	-	✓	✓	-	✓
Validation rule	-	✓	-	✓	✓	✓
Coversation rule	-	✓	-	-	✓	✓
History rule	✓	-	-	✓	-	✓

ingestion. These studies are recent work on ETL modeling and compared in Table 1.

2.2. Study on Text-Mining Techniques

Authors Schouten et al. [6], proposed Heracles, the model which is implemented and evaluate the text mining techniques, and taken considerations of a variety of mining software solutions in industry4.0. As compared to many text mining approaches, Heracles work on both implementations and evaluation timeline. The experiment results are proof that this model performs utmost in the mentioned domain.

This model [7] is provided the superficial knowledge of data mining which is used in text classification based on clusters, which is performing on how to classify the text, which gives more efforts on comparative analysis of different classification techniques on the text. Authors identified that none of any classification techniques perform utmost in all simulators on dataset considered by them in different classification approaches.

A US patent [8] says that the unveiled topic includes looking at the aftereffects of Natural Language Processing (NLP) of unstructured content to authentic outcomes for confirmation and approval of the NLP models/calculations. The investigation utilizes measurable hypotheses and practices to naturally screen and approves the exhibitions of the NLP calculations on an intermittent premise. Each unstructured content is gone through at least one NLP calculations and scored for importance or logical order. Conveyance of the scores is thought to be Gaussian in nature with the goal that likelihood esteem (p-esteem) might be created. At the point when the p-esteem is underneath edge esteem, manual labeling might be started for the present timespan to help retrain the models for better execution. Different epitomes are portrayed and guaranteed.

Authors Ran et al. [9], proposed a word division system dependent on knowledge word reference, and manufactured a word reference for insight data, which viably improved the exactness of word division in knowledge texts.

The motivation behind this investigation to the authors Rizum & Kucharska [10] is to create and to test a strategy that will distinguish chief purposes of clients' connections with style brands utilizing a lot of Text Mining Algorithms. The style business is one of the best in the online networking condition. A profound comprehension of design brand correspondence is intriguing from a hypothetical and useful perspective. The hypothetical estimation of this investigation adds to the internet based life brand information the board by giving a lot of picked up experiences because of the execution of the new methodological methodology introduced in this examination. The viable worth is the information about the nearness of design marks in web-based life got throughout the investigation.

An applied model is given by authors Wall & Singh [11] to clarify how ideas from pragmatics can improve existing content mining calculations to give increasingly exact data to dynamic. Switching the sober-minded procedure of significance articulation could prompt improved content mining calculations.

The authors Lamurias & Couto [12] presents an outline of the current biomedical content mining devices and bioinformatics applications utilizing them.

Authors Kowasari et al. [13], presents, a concise outline of content arrangement calculations is talked about. This diagram covers distinctive content component extractions, dimensionality decrease strategies, and existing calculations and procedures, and assessment techniques. At long last, the confinements of every strategy and their application in true issues are examined.

The authors Zaki & McColl [14], give a six-phase TMAR on the best way to utilize content mining techniques practically speaking. At each stage, the creators give a controlling inquiry, articulate the point, distinguish the scope of strategies, and show how AI and phonetic methods can be utilized. They find, At every one of the six phases, this paper exhibits helpful experiences that outcome from the content mining methods to give a top to bottom comprehension of the marvel and significant bits of knowledge for research and practice.

This investigation executed by Wang et al. [15], a coordinated arrangement of three characteristics—the

Table 2. Comparative Analysis of Different Text-mining models based on Data analysis, preperation and report functions.

Functions	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]
Universal-access	✓	✓	-	✓	✓	-	-	✓	-	-	✓	-	✓	✓
Analysis & Extraxtion	-	✓	-	✓	✓	✓	-	✓	-	-	✓	✓	✓	-
Cutom-dictionary	-	✓	-	-	✓	✓	-	-	-	-	✓	-	-	✓
Automatic-cleaning	✓	-	-	✓	-	-	-	✓	✓	-	✓	✓	✓	✓
Classification	-	✓	-	✓	✓	-	-	✓	-	-	✓	-	✓	✓
POS tagging	✓	✓	-	✓	✓	✓	✓	✓	-	-	✓	-	✓	-
Filtering	-	-	-	✓	-	-	-	-	✓	✓	-	-	-	✓
Concept-linking	-	-	✓	✓	-	-	✓	-	-	✓	-	-	✓	-
Word-segmentation	-	-	-	✓	-	✓	-	-	-	-	-	✓	-	✓
Text-clustering	-	-	✓	-	-	✓	✓	-	-	-	-	-	✓	-
Feature-extraction	-	-	-	-	-	✓	✓	-	-	-	✓	-	✓	✓
Text-absraction	✓	-	-	-	-	-	-	-	✓	-	✓	✓	-	-
Automatic-coding	✓	-	✓	-	-	-	-	✓	✓	✓	-	-	-	-
Reduction-technique	-	-	-	-	-	✓	-	-	-	✓	-	-	-	-
Natural-Lang-Query	-	-	-	-	-	-	-	-	-	✓	-	-	-	-
Indexing	✓	-	✓	-	-	✓	-	-	-	✓	-	-	-	✓
Similarity-search	-	-	-	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
Pinyin-search	-	-	-	-	-	-	✓	-	-	✓	-	✓	-	-
Proofreading	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Window-results	-	-	✓	-	✓	✓	-	✓	✓	✓	✓	-	✓	✓
Visualization-results	✓	✓	✓	-	✓	-	-	-	-	-	✓	-	-	✓
Vocabularies-sorting	-	-	-	-	✓	✓	-	✓	-	-	✓	-	✓	✓
Multilingual-support	✓	-	✓	✓	✓	✓	-	✓	-	-	✓	✓	-	-

#: ✓This checkmarks represnts the mentioned function is present.

Table 3. Transmission characteristics of recently emerged viruses.

Virus	Case fitality Rate(%)	Epidemic	Contained	Remars
COVID-19	unknown	✓	ǯ , efforts ongoing	
pH1N1	0.02-0.4	✓	ǯ , post epidemic circulation	
H7N9	39	ǯ	ǯ , erradiction efforts	
NL63	ǯ	ǯ	ǯ	
SARS-CoV	9.5	✓	ǯ	58% cases from transmission
MERS-CoV	34.4	ǯ	ǯ	70% cases from transmission
Ebola	63	ǯ	✓	70% cases from transmission

nearness of theme terms, number of non-point words, and proportion of the words with significant grammatical features—effectively affected various subjects. At last, we contrasted NoteSum and other existing synopsis frameworks. The outcomes showed that the NoteSum-produced outline was nearer to understudies' unique notes and subsequently brought about better execution in lucidness, usefulness, and fulfillment.

The authors Govindarajan et al. [16], presents a model to arrange a stroke that consolidates content mining instruments and AI calculations. AI can be depicted as a noteworthy tracker in regions like observation, medication, information the executives with the guide of appropriately prepared AI calculations. Information mining strategies applied in this work give a general audit about the following of data concerning semantic just as syntactic points of view.

In this paper, authors Pejic-Bach [17] build up a profile of Industry 4.0 employment commercials, utilizing content mining on openly accessible occupation promotions, which are frequently utilized as a channel for gathering important data about the necessary information and aptitudes in quick evolving ventures. We looked through the site, which distributes work promotions, identified with Industry 4.0.

Authors Sethi & Ramesh [18] presented three basic models for mining the text. The test results show that their proposed method mining calculations as far as runtime, memory utilization, join tallies, and

versatility.

Authors Thatha & Babu [19] proposed a simple to utilize structure to mine the text for highlight determination.

Authors Mourya & Kaur [20] diverse part deceives have applied with nonlinear classifiers for the characterization of content information mining. The consequences of proposed tests anticipate that help vector machine with outspread premise work accomplishes the most elevated in general precision.

Comparative Analysis of Different Text-mining models based on data analysis, preparation, and report functions is presented in Table 2.

2.3. Study on Virus

The target of study [21–23] recognizes regular reactions to the pandemic and how these reactions contrast across time. Also, experiences concerning how data and falsehood were transmitted using Twitter, beginning at the beginning times of this pandemic, are introduced. The bits of knowledge introduced in this work could help advise chiefs even with future pandemics, and the dataset acquainted can be utilized to obtain significant information to help alleviate the COVID-19 pandemic.

Authors Y Ji et al. [24] progressing plague of coronavirus illness 2019 (COVID-19) is crushing, despite broad usage in different kinds of measures. Where authors Mitja & Clotet [25] says that one of the fundamental suppositions of the model by Hellewell and partners is that all people with suggestive contamination with the extreme intense respiratory disorder (SARS) coronavirus 2 (SARS-CoV-2) are inevitably tried and detailed. In any case, under the rules of most nations with second rate transmission, clinicians will test speculated patients just on the off chance that they have gone to a scourge area since the episode started. A second presumption of the model is that seclusion of cases is 100% viable in halting transmission. However home control of tainted people and contacts is testing, viability is variable, and the thorough following included requires a lot of general wellbeing assets.

As per Wang et al. [26] the fast spread of new coronaviruses all through China and the world in 2019–2020 has greatly affected China's financial and social turn of events. Be that as it may, they additionally face the issues of graduated class' financial improvement challenges, the danger of savage contamination to clinical salvage groups and wellbeing laborers, disease of instructors and understudies, and the unacceptable use of data innovation in settling the emergency. Because of these dangers and crisis issues, we propose some comparing answers for open scattering, including issues identified with clinical security, crisis investigate, proficient help, positive correspondence, and various leveled data-based instructing.

This study is identified issues of background knowledge, pattern evaluation, transformation, and load deficiencies. This study proposed a novel transformation and join technique of dataset, with the minimum execution time and maximum accuracy benefits.

Transmission characteristics of recently emerged viruses are presented in Table 3.

3. Issues & Challenges

The outcome of state of art survey is issues and challenges in text mining methods are categorized into three-phase.

Methodology & user interaction. .

1. Mining different kinds of knowledge in the database;
2. Interactive mining of knowledge at multiple levels of abstraction;
3. Incorporation of background knowledge;
4. Data mining query languages and ad hoc data mining;
5. Presentations and visualization of data mining results;
6. Handling incomplete data;
7. Pattern evaluation

Performance Issues. .

1. Efficiency and scalability of data mining techniques;
2. Parallel, distributed, and incremental data mining algorithms

Diverse data type issue. .

1. Handling the relational and complex type of data;
2. Mining information from the heterogeneous database and global information system

Dataset Used. .

The dataset used in this research is also presented in Table 4. As per this table, it has 9 different datasets for a single record presentation. Each record is split into different tables. Once the record needs to be fetched is very difficult to view in a single window with all necessary information. In this dataset, each file has 8 to 11 attributes, for representing a record. This issue has been taken into consideration to solve through novel ETL techniques.

4. Methodology

The proposed work considered multiples "comma-separated values" CSV files from https://api.covid19india.org/csv/ and extract, transformation, and load ETL applied to process the files and get the knowledge from the dataset (COVID-19). The description of the dataset is presented in Table 4. The proposed model NETL is the inclusion of enrichment rule, selection criteria, coupling rule, validation rule, conversation rule, history rule, and apart from these cleaning, decision, and store function is used and compared in Table 2. The proposed work has four modules as mentioned in Figure 2.

1. Clustering
2. Text mining
3. ETL process
4. Data mining
5. Knowledge (Output)

4.1. Clustering

The datasets (Table 4) are processed through the K-nearest-neighbor KNN algorithm and classified into 4 clusters as awaited, positive, recovered, and dead. The objective of clustering is to classify the datasets into a relatively small number of classes that collectively classify and clustering on the similar data of the actual datasets.

4.2. Text mining

The meaningful information is extracting from the vast COVID-19 dataset. It focuses on identifying the different kinds of entities, columns (attributes), and the relationships among unstructured data.

4.3. A Novel ETL Approach

The proposed extract transform and load ETL method are presented in Figure 3.

To more clear about the data uploaded to the government website, the proposed model is based on

Table 4. Descriptions of dataset used

Dataset Description	Attributes											Total
	1	2	3	4	5	6	7	8	9	10	11	
raw_data1	city	district	state	code	notes	Ntl	type	status	src_1	src_2	src_3	18
raw_data2	Pnt_no	id	date	estd	age	gender	city	cntrd	status	src_1 & 2	src_3	21
raw_data3	enty	state	Pnt_no	date	age	gender	city	district	state	state_code	src_123	15
d_rec1	Date	Age	Gender	Pnt_Sts	City	District	State	code	Ntn	src_1	Src_2& 3	15
d_rec2	Date	Age	Gender	Pnt_Sts	City	District	State	code	Ntl	Src_1	Src_2& 3	10
st_wise	cnf	Rec	Death	act	Lst_Up	code	Cnf	Rec	Deaths	State_Notes		8
cs_tm_srs	Date	D_Cnf	Tot_Cnf	rec	Tot_rec	Dly_Dec	Tot_Dec	Deceased				9
dst_wise	Code	State	Key	District	Conf	Active	Recrd					31
state_wise	Date	Status	all state									

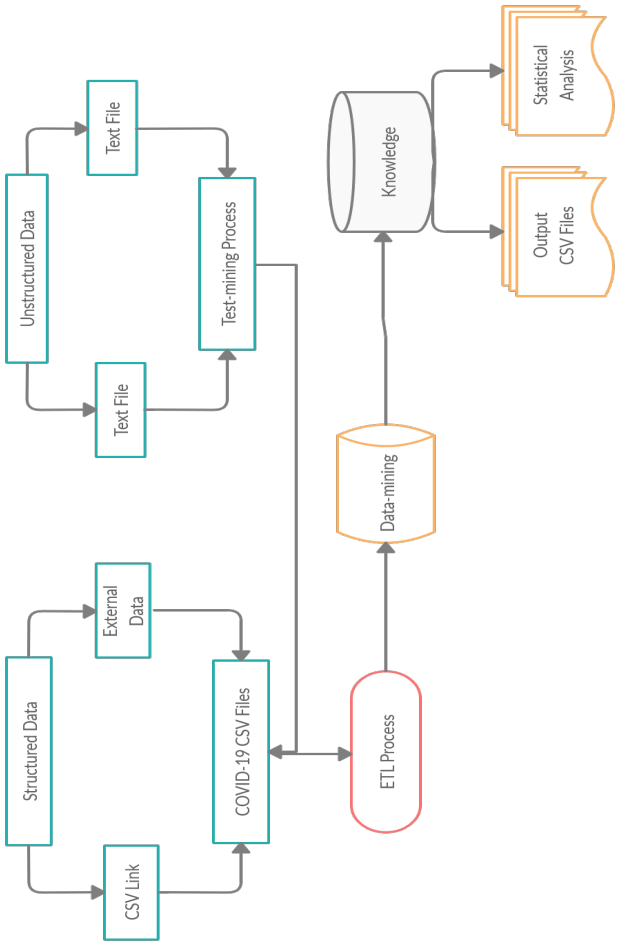

Figure 2. Research Methodology.

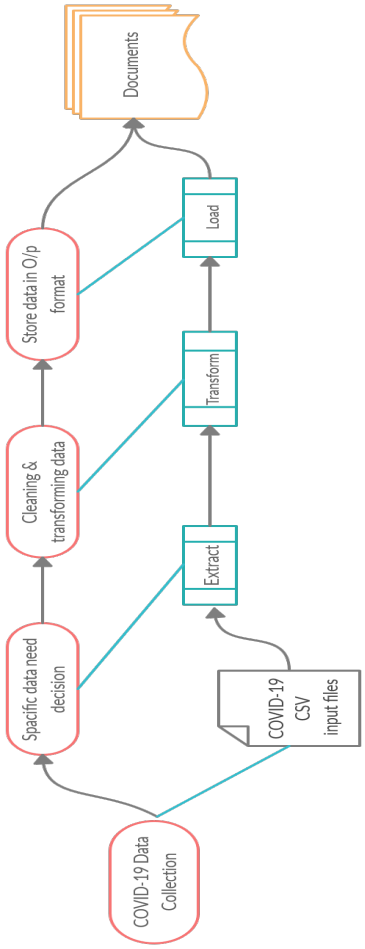

Figure 3. Preprocessing of data in ETL.

the ETL process, through which the uploaded data is extracted, clean, and then upload to the output files. A total of 9 CSV files (as described in Table 4) are processing in this module and a total of 110 attributes are identified. These attributes are processed and based on the relationship, they are transformed. The transformation process is presented in Figure 4.

The following steps are performed in the proposed ETL process:

Split. .
In this step, the input files are processed. The file is processed one by one and the number of rows is split into multiple rows (1:n). And the split rows have the information which is to seek to examine the number of positive cases in region-wise (refer Figure 5).

Validation. .
The missing information in rows is validating in this stage. If any field is left blank then it will verify the missing field value.

Join. .
All the rows which are split and verified through split and validation stages are joined through inner join rules.

5. Results & Analysis

The outcomes of the proposed novel ETL process are output file which has the description of total positive cases, active cases, recovery cases, and death rate, based on different regions. The steps discussed in the methodology section and its output are:

5.1. Splitting rows of input files

Splitting process inputs the CVS files (raw_data1, raw_data2, raw_data3) and produce the output file as presented in Table 5.

5.2. Joining

Joining process inputs the split files (refer Table 4) and produce the output file as presented in Table 6.

Table 5. Example of splitting rows of input files (raw_data1, raw_data2, raw_data3) to produce output file

Patient Number	State Pateint Number	Description						Source
		Detected city	District	State	Code	type	Test Status	
13699	KA-P361	Belagavi	Belagavi	Karnataka	KA	Hopitalized	Awaited	State
13699	KA-P361	Belagavi	Belagavi	Karnataka	KA	Hopitalized	Possitive	State
13699	KA-P361	Belagavi	Belagavi	Karnataka	KA	Hopitalized	Recovered	State
16246	TN-P1471	Viluppuram	Viluppuram	Tamil Nadu	TN	Hopitalized	Awaited	State
16246	TN-P1471	Viluppuram	Viluppuram	Tamil Nadu	TN	Hopitalized	Possitive	State
16246	TN-P1471	Viluppuram	Viluppuram	Tamil Nadu	TN	Hopitalized	Recovered	State
–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–
Total	52095	–	–	–	–	–	–	–

Table 6. Example of joining updated rows of input files (raw_data1, raw_data2, raw_data3) to procduce output file

Patient Number	State Pateint Number	Description						Source
		Detected city	District	State	Code	type	Test Status	
13699	KA-P361	Belagavi	Belagavi	Karnataka	KA	Hopitalized	Recovered	State
16246	TN-P1471	Viluppuram	Viluppuram	Tamil Nadu	TN	Hopitalized	Recovered	State
–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–
Total	17365	–	–	–	–	–	–	–

5.3. Analysis

The proposed novel ETL model is based on splitting rows, verifying the items (missing fields), and joining the split rows to find and convert the knowledge into the output file. The data set taken from the link `https://api.covid19india.org/csv/` is used to process in the proposed novel ETL. In input file total of 17365 rows are processed and split into 52095 rows and then inner join applied to get the output file. The output file shows only the results whether that patient is recovered instead of awaited or positive status.

Accuracy. .
Accuracy analysis has been done for proposed ETL model. It has been observed that while increasing the number of records, the proposed ETL model produces more accuracy as shown in Figure 6.

Failuare Count. .
The failure count is the total number of records fail (in percentage) has been calculated and shown in Figure 7. It has been observed that failure count is decreased while increases the number of records.

Execution Time Analysis. .
The execution time is recorded by the difference between submission time and completion time of the process. It is observed that the execution time for the proposed model is increasing while the number of records is increasing as shown in Figure 8.

The new findings of this research are, efficiently produce useful knowledge from the dataset as compilation time, failure count is less, and accuracy is high as compared to recent ETL research. The analysis is presented in Figure 9. The proposed NETL work processed records from 100 to 16000 and recorded accuracy, compilation time and failure count, and compared with BPMN [4] and it is observed that NETL is performing better than BPMN in above-given conditions.

The objective of this research is to provide a novel extract-transform-load technique to process the datasets(CSV files) and convert them into

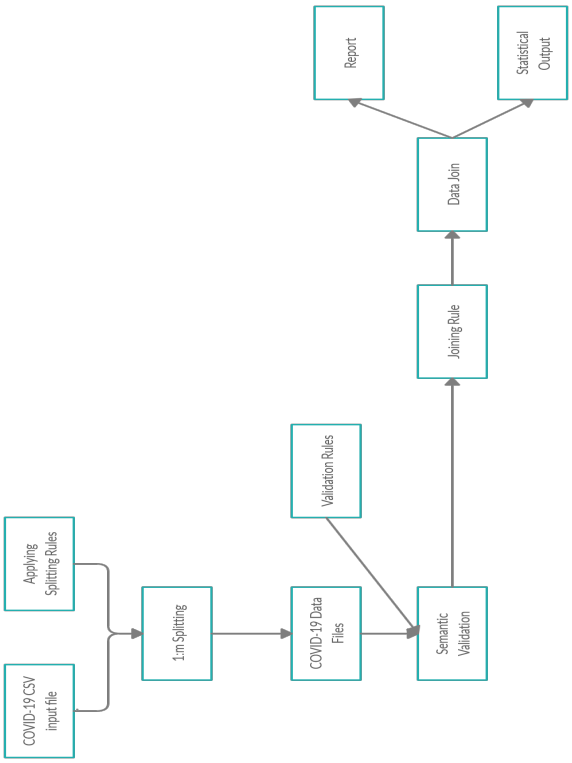

Figure 4. Details of Transformation step of proposed (ETL) approach which is applied to COVID-19 Data (CSV files)

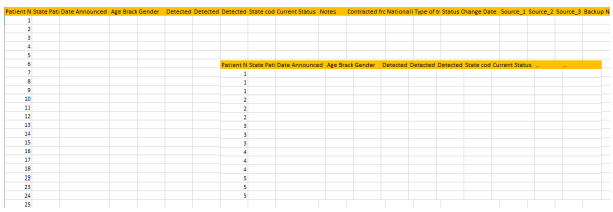

Figure 5. Example of splitting COVID-19 Data (CSV files)

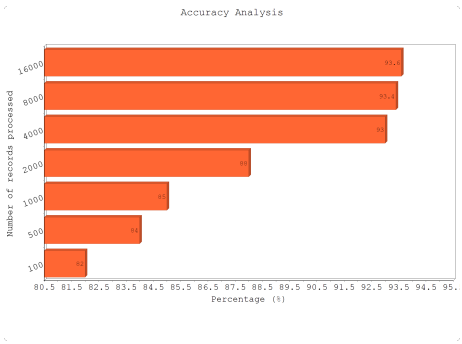

Figure 6. Accuracy Analysis of posposed ETL model while increasing number of records

meaningful knowledge to make a decision. The transformation, joining and loading is key to this

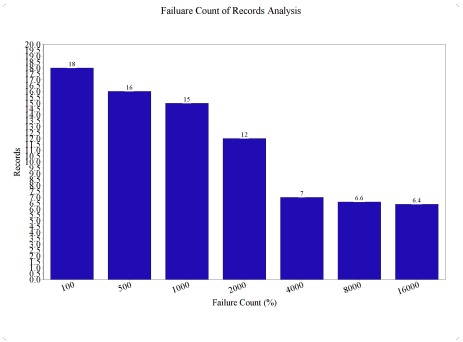

Figure 7. Failure count Analysis of posposed ETL model while increasing number of records

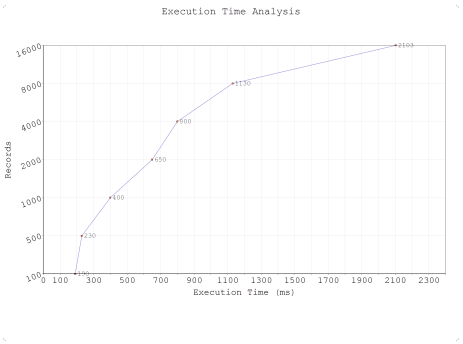

Figure 8. Excecution Time Analysis of posposed ETL model while increasing number of records

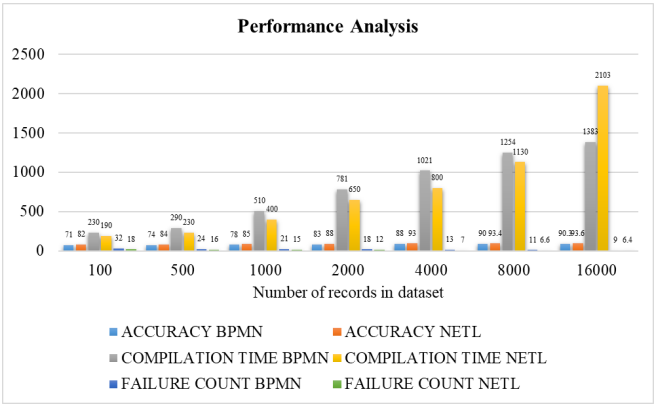

Figure 9. Performance Analysis

technique. The observations of the results are evidence that the proposed work gives better accuracy while increasing the size of the input file, and also the computational time is increasing but at the size of 16k, the computational time is saturated.

Moreover, the insight of this research work is to produce the knowledge from the multiple CSV files into single view form, authors recommendations to the readers are that they should use the current dataset, as results may slightly differ as datasets are updated

on daily basis from a source where the datasets are downloaded.

6. Conclusion

Coronavirus is declared a pandemic by world-health-organization WHO. To analyze the coronavirus data in India, a novel ETL (NETL) model is proposed. In this model, a total of 9 CSV files is processed as input files to get different results in different categories. The model designed to process any CSV file to produce the results. This model is having three modules namely splitting, verification, and join. The splitting module allows splitting the rows into three different rows for the used dataset, where the verification module is based on validation rules. The dataset is split into based on its coupling attributes and then joined with a single value to produce the updated results as per the current dataset. The last stage of this process is to join the data which is generated through splitting. The advantage of the proposed model is observed that it produces more accurate results while increasing the number of records and decrease failure count, but as records are increasing it takes more time to execute. A total of 17365 records are processed and it produces significant results.

NETL model is limited to 17365, as this model is tested in a static environment. In the future, the NETL model will access the dataset from the website directly through google co-lab, and any other platform to produce results as per the current time and updated dataset.

References

[1] Costello, T., & Blackshear, L. (2020). What Is ETL?. In Prepare Your Data for Tableau (pp. 1-3). Apress, Berkeley, CA.

[2] Galici, R., Ordile, L., Marchesi, M., Pinna, A., & Tonelli, R. (2020). Applying the ETL Process to Blockchain Data. Prospect and Findings. Information, 11(4), 204.

[3] Mallek, H., Ghozzi, F., & Gargouri, F. (2020). Towards Extract-Transform-Load Operations in a Big Data context. International Journal of Sociotechnology and Knowledge Development (IJSKD), 12(2), 77-95.

[4] Awiti, J., Vaisman, A. A., & Zimányi, E. (2020). Design and implementation of ETL processes using BPMN and relational algebra. Data & Knowledge Engineering, 101837.

[5] Semlali, B. E. B., El Amrani, C., & Ortiz, G. (2020). SAT-ETL-Integrator: an extract-transform-load software for satellite big data ingestion. Journal of Applied Remote Sensing, 14(1), 018501.

[6] Schouten, K., Frasincar, F., Dekker, R., & Riezebos, M. (2019). Heracles: A framework for developing and evaluating text mining algorithms. Expert Systems with Applications, 127, 68-84.

[7] Abikoye, O. C., Omokanye, S. O., & Aro, T. O. (2018). Text Classification Using Data Mining Techniques: A. Computing and Information Systems Journal, 1.

[8] Wang, H., Kuan, C. Y., Zhang, Y., & Zhao, R. (2018). U.S. Patent Application No. 15/586,739.

[9] Ran, Y., Liu, Q., Bo, Z., Li, Z., & Ke, D. (2018, May). Intelligence Information Retrieval Based on Text Mining. In 2018 7th International Conference on Energy, Environment and Sustainable Development (ICEESD 2018). Atlantis Press.

[10] Rizun, N., & Kucharska, W. (2018). Text Mining Algorithms for Extracting Brand Knowledge: The Fashion Industry Case. Available at SSRN 3148476.

[11] Wall, J. D., & Singh, R. (2017). Contextualized meaning extraction: A meta-algorithm for big data text mining with pragmatics. International Journal of Organizational and Collective Intelligence (IJOCI), 7(3), 15-29.

[12] Lamurias, A., & Couto, F. (2019). Text mining for bioinformatics using biomedical literature. Encyclopedia of bioinformatics and computational biology, 1, 602-611.

[13] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[14] Zaki, M. and McColl-Kennedy, J. (2020), "Text mining analysis roadmap (TMAR) for service research", Journal of Services Marketing, Vol. 34 No. 1, pp. 30-47.

[15] Wang, H. C., Chen, W. F., & Lin, C. Y. (2020). NoteSum: An integrated note summarization system by using text mining algorithms. Information Sciences, 513, 536-552.

[16] Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P., & Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms. Neural Computing and Applications, 32(3), 817-828.

[17] Pejic-Bach, M., Bertoncel, T., Meško, M., & Krstić, Ž. (2020). Text mining of industry 4.0 job advertisements. International Journal of Information Management, 50, 416-431.

[18] Sethi, K. K., & Ramesh, D. (2020). A fast high average-utility itemset mining with efficient tighter upper bounds and novel list structure. The Journal of Supercomputing, 1-31.

[19] Thatha, V. N., Babu, A. S., & Haritha, D. (2020). An Enhanced Feature Selection for Text Documents. In Smart Intelligent Computing and Applications (pp. 21-29). Springer, Singapore.

[20] Mourya, A. K., & Kaur, H. (2020). Performance and Evaluation of Different Kernels in Support Vector Machine for Text Mining. In Advances in Intelligent Computing and Communication (pp. 264-271). Springer, Singapore.

[21] Rodriguez-Morales, A. J., Cardona-Ospina, J. A., Gutiérrez-Ocampo, E., Villamizar-Peña, R., Holguin-Rivera, Y., Escalera-Antezana, J. P. & Paniz-Mondolfi, A. (2020). Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis. Travel medicine and infectious disease, 101623.

[22] Zhang, L., Jiang, Y., Wei, M., Cheng, B. H., Zhou, X. C., Li, J. & Hu, R. H. (2020). Analysis of the pregnancy outcomes in pregnant women with COVID-19 in Hubei Province.

Zhonghua fu chan ke za zhi, 55, E009-E009.
[23] Lopez, C. E., Vasu, M., & Gallemore, C. (2020). Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset. arXiv preprint arXiv:2003.10359.
[24] Ji, Y., Ma, Z., Peppelenbosch, M. P., & Pan, Q. (2020). Potential association between COVID-19 mortality and health-care resource availability. The Lancet Global Health, 8(4), e480.
[25] Mitja, O., & Clotet, B. (2020). Use of antiviral drugs to reduce COVID-19 transmission. The Lancet Global Health.
[26] Wang, C., Cheng, Z., Yue, X. G., & McAleer, M. (2020). Risk management of COVID-19 by universities in China.