A Profound Method for Three-Tier Toxic Word Classification using LSTM-RNN

Yogesh Devtulla
B.Tech III Year
School of Computing Science and Engineering
Galgotias University,
Greater Noida, India.
yogesh.devtulla@gmail.com

Sumit Baroniya
B.Tech III Year
School of Computing Science and Engineering
Galgotias University,
Greater Noida, India.
sumit.baroniya2002@gmail.com

Rishika Raj
B.Tech III Year
School of Computing Science and Engineering
Galgotias University,
Greater Noida, India.
rishikaraj287@gmail.com

N.Suresh Kumar
Assistant Professor
School of Computer Science and Engineering
Galgotias University,
Greater Noida, India.
sureshkumar0707@gmail.com

***Abstract*—Text classification is the activity of labeling natural language texts with relevant categories from a predefined set. In the research work, The words are categorized from a sentence according to their toxicity and according to modification will be done. As Online platforms such as social media provide opportunities to users to express their thoughts and their opinions on various topics and incidents. Due to the different opinion of every individual, an explicit language in some cases which hurt the readers. So, to solve a novel method is proposed by building a model wherein toxic words are categorized into three classes in keeping with their toxicity. In the primary class, words that are least toxic are indicated through crimson traces to the person, withinside the secondary class words are extra poisonous than the primary one, and again and again. The use of those phrases through individual, that words will be eliminated automatically and withinside the remaining or third class, words are the maximum toxic in assessment than that of the first and second class, through the use of those phrases' person account could be deleted/blocked. As a punishment, the account will be blocked for a specific time period and the length could be improved as in increasing of the occurrences of those words.**

Keywords— Toxic Phrases Classification, 3-Tier Toxic Word Classification, 3-Classes Toxic Word Classification, Machine Learning, Text Classification and Modification.

I. INTRODUCTION

The expansion of numerous social media platforms day by day. Has increasing as a result of User-generated content (UGC), which mainly refers to any type of information that users have created and made freely available online. It may contain a variety of information kinds, harsh language, hate speech, and cyberbullying. Due to the expansion of internet in every domain and all over the world which leads to the increment of the participation of people more actively and this creates a remark of their communication concern and their opinions in many forums. Although many of the comments are good and helpful but some of the comments are not good and even it hearts some other people. These abusive comments creates hatred feeling between the people. As these comments are public which means it can be read by anyone or it can be easily accessible. It is necessary to filter out that comments so as to stop hatred feelings among the people. The biggest issue for all researchers and developers is identifying these harmful comments. commenting on these harmful words today became a trend which has drawn a lot of interests not because people think commenting these will entertain and refrain people. And due to this many people start participating in this which diversely affects for all the creators. Although this field has made significant progress and uses a number of models, these models still have flaws and are unable to provide a precise answer to the issue at hand. The project's implementation as well as the relevant models were thoroughly described in this post.

The first step in developing appropriate counter measures for online toxic content would be to effectively recognize and monitor such information. The industry is currently looking at approaches that partially automate these operations in order to deal with Web size, despite the fact that social media organizations have been spending hundreds of billions of euros annually to manually filter and remove such information [15]. Differentiating between toxic [19] and non-toxic information and various toxic information categories is an enabling task for this approach. This task is referred as "toxic content identification" or "classification," broadly encompasses a wide spectrum of recent studies on identifying cyberbullying, abusive and insulting language, or hateful speech. Any biases or inaccurate connections in the training data can propagate unintentionally biased associations in the classification performance since algorithms for machine learning usually teach the easiest associations to predict the appropriate labels of inputs.

Basically, 3-tier system is built in which toxic words are classified into three categories according to their toxicity and according to that modification will be done. In the first category, words are least toxic which is indicated by red lines to the user, in the secondary category words are more toxic than the first one repeatedly using these words by a single user, that word will be removed automatically and in the last or third category, words are the most toxic [21] in comparison than that of first and second category, by using these words user account will be deleted and as a punishment, the platform will block these accounts for a short period of time and the period will be increased in the increases of the occurrences of these words.

II. LITERATURE SURVEY

Toxic comments refers to hatred online comments online social discussion platforms. Due to toxic comments many users are not able to put their points in online discussions. Here this paper by Ajay Kumar, Sharayu Lokhande, Kumar Shivam, Naresh Kumar, Rahul Malhan [1] they check the toxicity of comment. And if the comment is toxic then they classify the comments into different categories to examine the type of toxicity. Maintaining the Integrity of the Specifications The problem of classification is being addressed in the data mining, machine learning, database, and information retrieval communities, with applications in various domains such as: marketing, medical diagnostics, newsgroup filtering, and document organization are extensively researched. This post provides an overview of various text classification algorithms [2].

In this work, they combine the strengths of both architectures and propose a novel and unified model called C-LSTM (Convolutional Neural Networks based Long Short-Term Memory) [3] for sentence representation and text classification. This paper says that C-LSTM uses CNN (Convolutional Neural Networks) [17] [23] to extract a sequence of higher-level phrase and are input into a LSTM-RNN (Long Short-Term Memory Recurrent Neural Network) to obtain the sentence representations. Sites fail to properly promote discussions, which forces many communities to restrict or disable user comments. In order to analyse the toxicity as accurately as possible, this paper will rigorously assess the prevalence of online harassment [4].

With so much content being shared by users of social media and other websites every second, it becomes important for those platforms to spot hazardous content. Traditional algorithms that rely on users reporting harmful content for it to be removed and required steps to be made against the users posting the content would take a lengthy period, during which it would have attracted media attention and resulted in significant arguments over the content [5]. The classification of toxic comments has become into an active research area with numerous recently proposed methodologies. In order to do this, Betty van Aken, Julian Risch, Ralf Krestel and Alexander Löser [6] test various deep learning [25] [26] and shallow approaches on a fresh, sizable dataset of comments and suggest an ensemble that beats all individual models. They also confirm the results using a different dataset. The ensemble's results allow us to do a thorough error analysis, which identifies unresolved issues with current techniques and suggests areas for ongoing future research. These difficulties include inaccurate dataset labelling and a lack of paradigmatic context.

Text classification has become a crucial task to preserve the data available in a systematic manner since textual material online is expanding quickly. Here in this paper, they used these two extensively used methods are Word2vec and TF-IDF [7]. Quora is an online community where users may ask questions and share helpful information to learn from one another's knowledge. They face a significant task in weeding out queries with such phoney motives, such as those that promote hatred, make generalisations, etc. On social media platforms, arguments frequently break out during conversations and debates and entail the use of poisonous comments, which are nasty, insulting, and hateful remarks. Regression vector voting classifier (RVVC) [8], a collective methodology, is introduced in this paper as a method for locating harmful comments on social media platforms. On the imbalanced and balanced dataset, a number of experiments are run to evaluate the effectiveness of the suggested methodology.

In this work, text classification is taken as the principle focus, an unique and adaptable model that uses a Recurrent Neural Network [9] and Capsule Network captures more contextual information when learning word representations in the text. On Wikipedia`s talk page changes provided by Jigsaw in Kaggle's hazardous comment classification, a number of tests are carried out. The method most frequently employed for conducting a quantitative study of remote sensing picture data is supervised classification [10]. The fundamental idea behind it is to divide the spectral domain into regions that can be connected to the ground cover classes that are relevant to a given application. For the purpose, a number of algorithms are available; the most popular ones are covered in this chapter. Professional-looking documents like applications, forms, templates, business cards, letters, paper, reports, and booklets can be created, edited, printed, and shared using Microsoft Word. The word processing program Word was created by Microsoft, a global technology business based in the United States [11] [18].

Social network data mining is a method for determining public sentiment that is both efficient and reliable. To provide a platform for social network analysis, a sentiment analysis program is being developed. In this study, 3000 Reddit data and 3000 Twitter data were collected, organized, examined, and graphically represented[12]. 83% and 77%, respectively, are great percentage results for the Twitter and Reddit data. A machine learning model is developed iteratively. A data scientist cannot effectively manage models that are created over time. This work is processed using ModelDB, a cutting-edge end-to-end system for managing machine learning models. It is explained by the authors Manasi Vartak, Wei-En Lee, Srinidhi Viswanathan[13]. Through a web-based interface, the ModelDB frontend enables visual model exploration and analysis.

Predicting textual article classifications is the aim of automatic text categorization [22], particularly in the medical industry. Some applications require the data to be used to be innately defined by more than one label. When compared to analogous efforts already published in the literature, the experiment of the approach, known as GL-LSTM[14].Based on an assumed cardiovascular text dataset, has yielded excellent results with an overall accuracy of 0.927.

III. METHODOLOGY

To identify toxicity and accuracy of toxicity in public remarks posted on various forums and social media platforms this research utilizes a variety of text classifiers. The datasets used to train these classifiers determine how they are trained. A used training dataset consists of 159571 comments that have been classified as poisonous and neutral. Toxic terms are labelled in three categories: "Primary," "Secondary," and "Tertiary." LSTM is one type of RNN Technique. It is mostly employed in situations involving sequence predictions and used handle the order dependence in a sequence. This approach is frequently used to solve text classification [21] issues and aids in preserving the word order that can affect a document's meaning.

Recurrent neural networks will be used to analyses sequence data, which can include time series data, audio data, and sentences of text. In general, sequence refers to data that is presented in a specific order and where the order affects the information contained in the data. Therefore, before feeding it to the model, the order is crucial. Because there is no way for standard neural networks to recall prior outputs, they are unable to handle sequence data. RNN solve this problem by simply using its own output as an input, preserving the information.

The first step needed to prepare data for this technique purpose of LSTM is to group the words together and represent them as integer values, where words that are unique are represented by integers. In order to incorporate it in the model, vocabulary size is defined, which determines the overall amount of frequently occurring terms in text. Now I was able to utilize two different deep learning models in this research and apply them to a use-case for natural language processing. I became aware of effective ways to clean text information due to the project's multiple data pre-processing and feature engineering stages. I was able to comprehend how different deep-learning models, including CNN, LSTM, and the LSTM-CNN hybrid model, worked. I learned about word embedding concepts and the benefits of using pre-trained word embedding.

IV. Proposed System

Initially the data are collected and retrieve all the element of the dataset by using import command to use dataset in the program. By using the info command imported dataset is described using which the model get to know, number of rows and columns, name of column, what kind of dataset they contain, does they contain null values or not. Furthermore, Data preprocessing transforms raw data into a format that computers and machine learning algorithms can comprehend and evaluate. It is a step in the data mining and data analysis process. In this work, 3 phases of novel method is proposed for toxic word classification.

Firstly, Evolution of Data quality /Checking for missing values. using the "isnull" function on both the training and test data. The model is ensures, no missing records, therefore it continues with data cleaning/Text Normalization will be follow up like this Eliminating spaces between text characters, Eliminating Repeated Characters, lowering the case of data, Elimination Punctuation, Removing unnecessary spaces between words. Then moving ahead towards Data transformation/Lemmatization using developed the function "lemma" to do lemmatization on the clean data and imported "WordNetLemmatizer" from the "nltk" library. Lastly in data preprocessing Data compression it is the science and practice of storing information in a compact way is known as data compression. There are numerous compression packages used to compress files, as one would have realized. Algorithm speedups, storage cost savings, and cost-effective transmission are all benefits of compression.it also include Tokenization, Indexing, Index Representation, Padding.

The data is preprocessed and cleaned for building the model. Using the machine learning, the model is categorized as train-set and validation-set. It is appropriate to adopt the Long Short Term Memory model (LSTM) on Natural Language Processing use-case. The main distinction between LSTM networks [16] and RNNs is the replacement of hidden layer updates with memory cells. They become more adept at spotting and exposing distant data relationships, which is crucial for sentence structures. Now, model is created and passing on to the training and testing of the data models. As compared to CNN, which can extract the local properties of the text, LSTM can properly preserve the features of past data in extended text sequences. So I determined the LSTM model's necessary number of layers, assembled the model, and then trained it with the train dataset. Followed with the Observe Evaluation Metrics I got encouraging results when I evaluated the model based on accuracy and loss numbers. It gave me the assurance that I needed to use the testing set to evaluate how well my deep learning models performed. Now discussing and analyzing, the results are analysed.

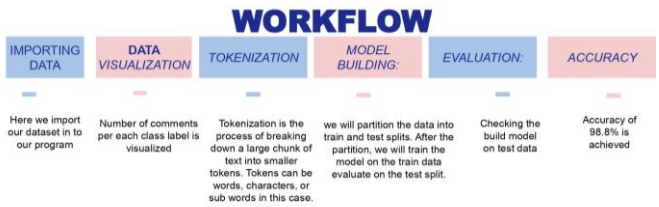

Fig. 1. Workflow of the model

Fig.1. shows the guidelines which was carried out in this research work. Initially, all necessary libraries and datasets are imported. In the data visulization phase, the comments were added to each class labels and the data are categorized for displaying. In the next phase of tokenization, train and test portion of the data is prepared for building the model. The model is trained using the GPU and achieved the accuracy rate of 98.8%.

V. Result and Outcomes

An encouraging result are obtained, when the model is evaluated based on accuracy and loss. It gave me the assurance I needed to use the test set to evaluate how well my deep learning models performed.

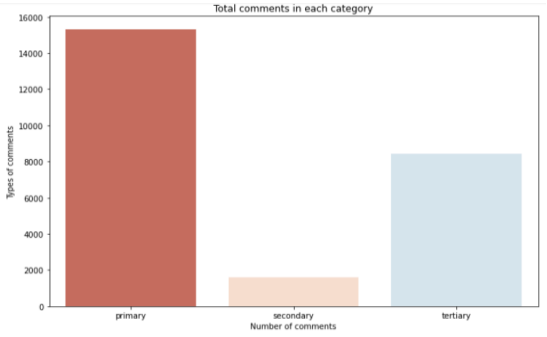

Fig. 2. No. of comment in each category

Here in Fig.2. shows types of comments in a dataset used and total comments in each category i.e. primary, secondary and tertiary. Reddish brown bar chart is showing 15000+ comments in primary, light pink bar chart is showing 1000+ comments in secondary and light blue bar chart is showing 8000+ comments in tertiary section.

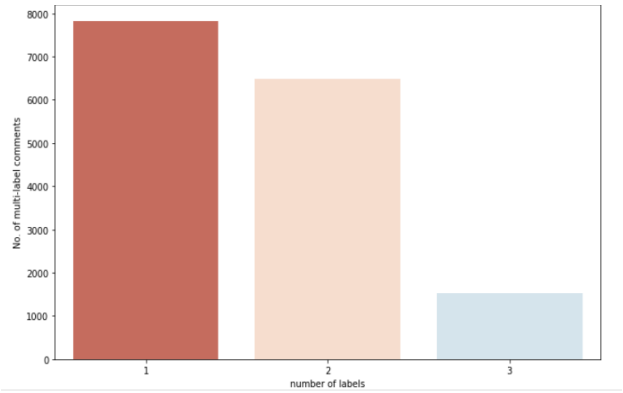

Fig. 3. Number of labels in a dataset

Fig.3. shows number of labels in a dataset and number of multi-label comments are used in a label. Reddish brown bar chart is showing 7500+ comments in label 1, light pink bar chart is showing 6000+ comments in label 2 and light blue bar chart is showing 100+ comments in label 3.

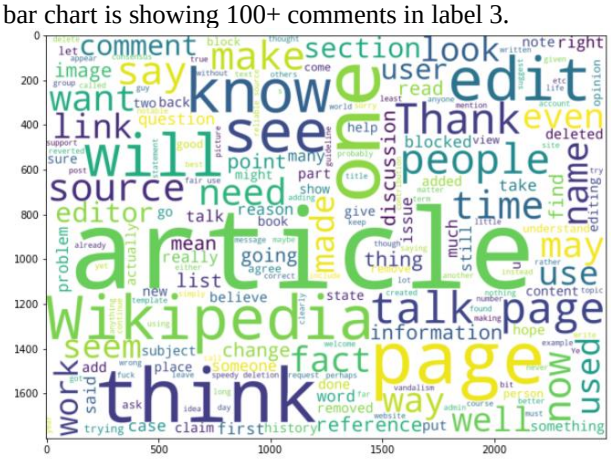

Fig. 4. Frequent words used in the text

Fig.4. shows the frequent words used in the text, world cloud is used for this purpose which gives an understanding of what kind of words are being used. The frequent words seen from the word cloud currently does not make any meaningful difference [24]. It can easily be seen that words are more related to online discussions.

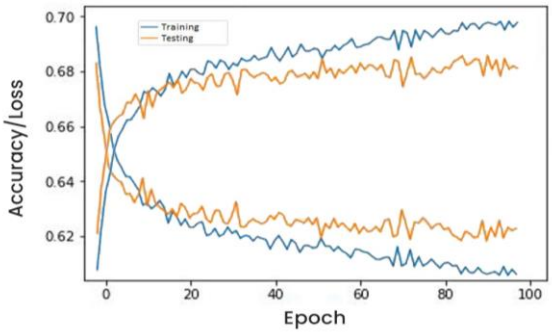

Fig. 5. Accuracy and hamming loss of the model

This Fig.5. Shows the Accuracy score and Hamming Loss. Hamming Loss and Accuracy Score for all Three Labels in the Output are recorded.

VI. CONCLUSION

This research aims to analyze classification methods utilized in toxic text classification as well as research on LSTM technique. For the classification of massive amounts of data, the field of toxic text classification integrates methods from several other areas, including as machine learning, artificial intelligence, pattern recognition, and many more. Classification approaches are frequently effective at modelling interactions. Depending on the circumstance, one of these methods may be more appropriate than the others. Here, the data are divided into three categories—primary, secondary, and tertiary—based on their levels of toxicity, and the necessary adjustments are made in accordance to the model. These classification algorithms can be used with data from several social media platforms, including Facebook, Instagram, YouTube, Twitter, and others. "Facebook has 2.4 billion members worldwide, making it the most popular social media platform. More than one billion people utilize WhatsApp and YouTube, two additional social media sites." Due to the enormous number of users on these social media platforms, a significant amount of data is generated every day, and special techniques are needed to identify material that is hazardous. Therefore, utilizing LSTM classification approaches, the model has a 98.8 percent accuracy rate.

As a result, the work suggests an effective sequence model architecture that may be applied to numerous applications that call for the assessment of statement toxicity. As a result, figuring out how harmful each remark is takes time and effort. Any application will see a significant reduction in this, which can avert conflicts and promote social peace. The online platform's users help the platform accomplish its goal. They are accused of giving access to harmful content on the platform. This will greatly minimize the time and effort needed to assess the toxicity of each comment in applications of any kind, allowing disputes to be avoided and preserving a pleasant environment among the participants. In future, the work can be optimized by increasing the accuracy rate and reduce the time of processing. It can also be expanded to all the social media platforms to detect and eradicate the toxic words.

VII. REFERENCES

[1] Ajay Kumar , Sharayu Lokhande , Kumar Shivam , Naresh Kumar, Rahul Malhan, 2021, Multilabel Toxic Comment Detection and Classification, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 05 (May 2021).

[2] Aggarwal, C.C. and Zhai, C.X. (2012) “A survey of text classification algorithms,” Mining Text Data, pp. 163–222. Available at: https://doi.org/10.1007/978-1-4614-3223-4_6.

[3] Chunting Zhou, Chonglin Sun, Zhiyuan Liu and C.M. Lau. (2015) “A C-LSTM Neural Network for Text Classification survey of text classification algorithms,”. Available at: https://doi.org/10.48550/arXiv.1511.08630.

[4] Rahul, Kajla, H., Hooda, J., & Saini, G. (2020). Classification of online toxic comments using machine learning algorithms. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). https://doi.org/10.1109/iciccs48265.2020.9120939

[5] Greevy, E. and Smeaton, A.F. (2004) “Classifying racist texts using a support vector machine,” Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04 [Preprint]. Available at: https://doi.org/10.1145/1008992.1009074.

[6] Betty van Aken, Julian Risch, Ralf Krestel, Alexander Löser. (2018) “Challenges for Toxic Comment Classification,” An In-Depth Error Analysis. Available at: https://doi.org/10.48550/arXiv.1809.07572.

[7] MZhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics, 1(1-4), 43–52. https://doi.org/10.1007/s13042-010-0001-0.

[8] Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Impact of smote on imbalanced text features for toxic comments classification using RVVC model. IEEE Access, 9, 78621–78634. https://doi.org/10.1109/access.2021.3083638 hang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics, 1(1-4), 43–52. https://doi.org/10.1007/s13042-010-0001-0

[9] Deshmukh, S., & Rade, R. (2018). Tackling toxic online communication with recurrent capsule networks. *2018 Conference on Information and Communication Technology (CICT)*. https://doi.org/10.1109/infocomtech.2018.8722433

[10] Richards, J. (2012). Supervised classification techniques. Remote Sensing Digital Image Analysis, 247–318. https://doi.org/10.1007/978-3-642-30062-2_8

[11] (n.d.). Supplemental Information 1: Microsoft Word - Supplemental Table S1.Doc Table S1. https://doi.org/10.7287/peerj.preprints.2303/supp-1

[12] Social network analysis using Python Data Mining. IEEE Xplore. (n.d.). Retrieved November 8, 2022, from https://ieeexplore.ieee.org/document/9268866

[13] MIT, M. V., Vartak, M., Mit, Profile, M. I. T. V., MIT, H. S., Subramanyam, H., MIT, W.-E. L., Lee, W.-E., MIT, S. V., Viswanathan, S., MIT, S. H., Husnoo, S., MIT, S. M., Madden, S., MIT, M. Z., Zaharia, M., & Metrics, O. M. V. A. (2016, June 1). ModelDB: Proceedings of the workshop on human-in-the-loop data analytics. ACM Other conferences. Retrieved November 8, 2022, from https://dl.acm.org/doi/10.1145/2939502.2939516

[14] R. Chaib, N. Azizi, N. E. Hammami, I. Gasmi, D. Schwab and A. Chaib, "GL-LSTM Model For Multi-label Text Classification Of Cardiovascular Disease Reports," 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), 2022, pp. 1-6, doi: 10.1109/IRASET52964.2022.9738147.

[15] Ortiz-Ospina, E. (2019) The rise of Social Media, Our World in Data. Available at: https://ourworldindata.org/rise-of-social-media (Accessed: November 8, 2022).

[16] Gangwar, S., Bali, V. and Kumar, A. (2018) “Comparative analysis of wind speed forecasting using LSTM and SVM,” ICST Transactions on Scalable Information Systems, p. 159407. Available at: https://doi.org/10.4108/eai.13-7-2018.159407.

[17] Georgakopoulos, S.V. et al. (2018) “Convolutional Neural Networks for toxic comment classification,” Proceedings of the 10th Hellenic Conference on Artificial Intelligence [Preprint]. Available at: https://doi.org/10.1145/3200947.3208069.

[18] S. Z. Mishu and S. M. Rafiuddin, "Performance analysis of supervised machine learning algorithms for text classification," 2016 19th International Conference on Computer and Information Technology (ICCIT), 2016, pp. 409-413, doi: 10.1109/ICCITECHN.2016.7860233.

[19] M. Husnain, A. Khalid and N. Shafi, "A Novel Preprocessing Technique for Toxic Comment Classification," 2021 International Conference on Artificial Intelligence (ICAI), 2021, pp. 22-27, doi: 10.1109/ICAI52203.2021.9445252.

[20] S. Deshmukh and R. Rade, "Tackling Toxic Online Communication with Recurrent Capsule Networks," 2018 Conference on Information and Communication Technology (CICT), 2018, pp. 1-7, doi: 10.1109/INFOCOMTECH.2018.8722433.

[21] V. Swetha, R. Anuhya, E. S. Sowmya and A. Geethanjali, "Building a Toxic Comments Classification Model," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2021, pp. 1519-1523, doi: 10.1109/ICECA52323.2021.9675911.

[22] S. Jain, G. Kaushik, P. Prabhu and A. Godbole, "Detox: NLP Based Classification And Euphemistic Text Substitution For Toxic Comments," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-5, doi: 10.1109/ICCCNT51525.2021.9579846.

[23] N. S. Kumar, A. K. Goel and S. Jayanthi, "A Scrupulous Approach to Perform Classification and Detection of Fetal Brain using Darknet YOLO v4," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 578-581, doi: 10.1109/ICACITE51222.2021.9404656.

[24] N. S. Kumar and A. Kumar Goel, "An Optimized Approach to Clinical Object Identification using YOLO v3 in the Cloud Environment," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 856-859, doi: 10.1109/ICACITE51222.2021.9404592.

[25] A. kumar, A. Kumar, P. Kumari and N. Suresh kumar, "A Pragmatic Approach to Face Recognition using a Novel Deep Learning Algorithm," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 806-810, doi: 10.1109/ICACITE51222.2021.9404697.

[26] N. Suresh Kumar, and Amit Kumar Goel. Detection, Localization and Classification of Fetal Brain Abnormalities using YOLO v4 Architecture [J]. Int J Performability Eng, 2022, 18(10): 720-729.