Classification Of Online Toxic Comments Using Machine Learning

G.Sreelekha1, V.Kousalya[2]

[1]AssistantProfessor, Dept of MCA,Audisankara College of Engineering &Technology (AUTONOMOUS), Gudur , Tirupati (Dt), AP, India.

[2]PG Scholar, Dept of MCA, AudisankaraCollege Of Engineering &Technology (AUTONOMOUS) Gudur , Tirupati (Dt), AP,Indian.

ABSTRACT

Toxic comments are disrespectful, toxic, abusive and unreasonable that makes the users leaves the discussion. In present generation, social media has become major part of everyone's life. People are getting bullied for numerous reasons. Not all people on internet are interested in participating nicely, some will vent their anger, insecurities and prejudices. This anti-social behaviour often occurs during the debates in comment section, discussion and fights often takes place in the online platform where it involves rude and disrespectful comments which are known are toxic comments. Comments containing explicit language can be classified into myriad categories such as Toxic, Severe Toxic, Obscene, Threat, Insult, and Identity Hate. The threat of abuse and harassment means that many people stop expressing themselves and give up on seeking different opinions. To protect users from being exposed to offensive language on online forums or social media sites, companies have started flagging comments and blocking users who are found guilty of using unpleasant language. Several Machine Learning models have been developed and deployed to filter out the unruly language and protect internet users from becoming victims of online harassment and cyberbullying We will aim to create a classifier which classifies the comments between the toxic and non- toxic comments which helps the organizations to get the better picture of the comment section to examine the toxicity with high accuracy using Lstm-cnn model.

I. INTRODUCTION

Social media is a place where a lot of discussions happen, being anonymous while doing so has given the freedom to many people to express their opinions freely. But people who disagree with a point of view extremely can misuse this freedom sometimes. Sharing things that you care about will become a difficult task with this constant threat ofharassment or toxic comments online. This will eventually lead to people not sharing their ideas online and stop asking for other people's opinion on them. Unfortunately, the socialmedia platforms face these issues all the time and find it difficult to identify and stop these toxic remarks before it leads to the abrupt end of conversations.

In this we will be using Natural Language Processing with Deep neural networks to solve this problem of identifying the toxicity of online comments. Word embeddings will be used in conjunction with recurrent neural networks with Long Short Term Memory (LSTM), Convolutional Neural Networks (CNN), and separately and see which model fits and works best.Text classification has become one of the most useful applications of Deep Learning, this process includes techniques like Tokenizing, Stemming, and Embedding. This paper uses these techniques along with few algorithms,that are

used to classify online comments based on their level of toxicity.

We proposed a neural network model to classify the comments and compared the model's accuracy with some other models like Long Short Term Memory (LSTM) and Convolutional Neural Network .The comments are first passed to a tokenizer or vectorizer to create a dictionary of words, then an embedding matrix is created after which it is passed to a model to classify.

II. LITERATURE REVIEW

Toxic comments on social media platforms have been a source of a great stir between individuals and groups. A toxic comment is not only verbal violence but includes the comment that is rude, disrespectful, negative online behaviour, or other similar attitudes that make someone leavea discussion. Therefore, the toxic comments identification on social platforms is an important task that can help to maintain its interruption and hatred-free operations. Consequently, a large variety of toxic comment approaches have been proposed. Three characteristics concerning toxic classification are evaluated: classification, feature dimension reduction, and feature importance.

The author of "Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification using RVVC Model" discusses about the ensemble approach calledregression vector voting classifier(RVVC), to identify the toxic comments over the social media platforms. The dataset for this is taken from Kaggle which is a multi-label dataset which contains labels as toxic, severe_toxic, threat, insult, and identity_hate. The values of the dataset are given as binary values which contains 158,640 comments in total with toxic comments. Owing to the study of higher accuracy ensemble model their experiment indicated the good performance and they combined the LR and SVC model to get the higher accuracy which is known as RVVC. They conducted the experiments using originally imbalanced dataset with TF-IDF and BoW separately. The results proved that RVVC outperforms all other individual models when TF-IDF features are used with SMOTE balanced dataset and achieves an accuracy of 0.97. Despite the better performance of proposed model, its computational complexity is higher than individual models. The authors for "Toxic Comment Classification Implementing CNN combining Word Embedding Technique" says that Despite this model, it only classifies into six labels which further improves to another model where it first classifies whether the comment is positive or negative and then classifies into six labels. The dataset used is Wikipedia's talk page edits, collected from Kaggle. They proposed the ensembled model called Convolution Neural Network(CNN) which is structured as data cleaning, adopting the NLP techniques, stemming, and converted word into vector by word embedding techniques. The accuracy for this model is calculated based on ROC – AUC. The receiver operating characteristic curve (ROC) score is 98.46% and the area under score (AUC) score for this model is 98.05% which is much accurate than the previous model and the existing works. There's another technique where it uses the deep learning model recurrent neural networks(RNN) to perform text classification on a multi- label text dataset to identify different forms of internet toxicity. The paper "Application of Recurrent Neural Networks in Toxic Comment Classification" discusses about this aspect. The dataset used for this methodology is the public dataset which is provided by Conversation AI team. Even Though the previous models has given the accurate toxicity scores, the models still miss classify some texts that share similar patterns as toxic comments which can be reduced using the RNN method. They used the Word2Vec embedding model to train the model to remove the noise in the data and the methodologies used

are the recurrent neural network and done the comparison analysis with GRU. They have successfully employed the word2vec embedding and recurrent neural network in building a toxic comment classification model and achieved high accuracy with low cost. Even with the existing models have achieved high accuracy, building the model takes more time and complexity of the model will higher. This leads to Automatic detection. The paper "Automatic toxic Comment Detection Using DNN" discusses about the automatic tools which where buildusing the LSTM and RNN to improve the accuracy and provide the results much faster than the traditional methods. This paper compared three state-of-art unsupervised word embedding models which were Mikolov's word embedding, fastTextsubword embedding, BERT Wordpiece model. The dataset used is Wikipedia Detox Corpus which talks about English Wikipedia talk pages. All the models were compared against the BERT fine-tuning. And on experiments and results, it showed that BERT fine-tuning is the most efficient model at this automatic toxic classification task. This can further developed into the speech toxic recognition using more advanced models like XLNet model. There's another model where the hybrid models are produced to get the accurate score much more than CNN model. The paper "Toxic Comment Classification Using Hybrid Deep Learning Model" discusses about the hybrid models used which are Bidirectional gated recurrent network, convolution neural network and they achieved the accuracy of 98.39% and the f1 score of this model was 79.91%. As much as toxic comment classification is important, toxic span prediction also play the similar role which helps to build more automated moderation systems. The paper "Multi-task learning for toxic comment classification and rationale extraction" discusses about the multi-task learning model using the Toxic XLMR for bidirectional contextual embeddings of input text for toxic comment classification and a Bi-LSTM CRF layer for toxic span and rationale identification. The dataset used was curated from Jigsaw and Toxic span prediction dataset. The model has outperformed the single task models on the curated and toxic span prediction models by 4% and 2% improvement for classification. The future improvements can be added totake more delicate context and handle the subtle differences in usage of keywords. The paper "Vulgarity Classificationin comments using SVM and LSTM" discusses about the hybrid model of SVM and RNN-LSTM. The data is vectorized using TF-IDF and bag of words. This paper also discusses the nature of the dataset. The results found to give a promising assurance in finding a solution. The paper "Modern Approaches to Detecting and classifying Toxic Comments using Neural Networks" discusses about the algorithms constructed using deep learning technologies and neural networks that solve the problem of detecting and classifying toxic comments. The algorithms are tested and trained on a large training set and tagged by Google and Jigsaw which was taken from Kaggle.

III. PROPOSED WORK

A. DATA DESCRIPTION

This study aims at the automatic classification of toxic and non-toxic comments from social media platforms. Various machine learning models are utilized for this purpose to evaluate their strength for the said task. For evaluation, the selected models are trained and tested with binary class datasets. Traditionally, toxic comments are grouped under several classes such as hate, toxic, threat, severe toxic, obscene, insult and non-toxic, etc. We follow a different approach by grouping

the comments under two classes, toxic and non-toxic. The original dataset which is taken from Kaggle [30], is a multi-label dataset and contains labels such as toxic, severe_toxic, obscene, threat, insult, and identity_hate. The non-toxic comments belong to one class, while from the other comments only those comments are selected that have toxic labels. It means that the comments that label severe_toxic, obscene, threat, insult, and identity_hate are not selected.

B. PREPROCESSING STEPS

Pre-processing techniques are applied to clean the data which helps to improve the learning efficiency of machine learning models [31]. For this purpose, the following steps are executed in the given sequence.

Tokenization: is a process of dividing a text into smaller units called 'tokens'. A token can be a number, word, or any type of symbol that contains all the important information about the data without conceding its security.

Punctuation removal: involves removing the punctuation from comments using natural language processing techniques. Punctuations are the symbols that are utilized in sentences/comments to make the sentence clear and readable for humans. However, it creates problems in the learning process of machine learning algorithms and needs to be removed to improve their learning process. Some common punctuation marks are mostly used such that colon, question marks, comma, semicolon, full-stop/period, etc. ?:,;.[]() [32].

Number removal: is also a part of preprocessing which helps to improve the performance of the machine learning algorithms. Numbers are unnecessary and do not contribute to the learning of text analysis approaches. Removing the numbers increases the efficiency of models and decreases the complexity of the data.

Stemming: is an important part of preprocessing because it increases the performance by clarifying affixes from sentences/comments and converting the comments into the original form. Stemming is the process of transforming a word into its root form. For example, different words have the same meaning such as: 'plays', 'playing', 'played' are modified forms of 'play'. Stemming is implemented using the Porter stemmer algorithms [33].

Spelling correction: is the process of correcting the misspelled words. In this phase, the spelling checker is used to check the misspelled words and replace them with the correct word. Python library 'pyspellchecker' provides the necessary features to check the misspelled words and is used for the experiments [34].

Stopwords removal: Stopwords are those English words that do not add any meaning to a sentence. So these can be removed by stopwords removal without affecting the meaning of a sentence. The removal of stop-words increases the model's performances and decreases the complexity of input features [35].

C. FEATURE ENGINEERING

Feature engineering aims at discovering useful data features or constructing features from original features to train machine learning algorithms effectively [36]. The study [37] concludes that feature engineering can improve the efficiency of machine learning algorithms. 'Garbage out' is a corporate proverb used in machine learning which implies that senseless data used as the input, yields meaningless output. In contrast, more information-driven data will yield favorable results. Hence, feature engineering can derive useful features from raw

data which helps to improve thereliability and accurateness of learning algorithms. In the proposed methodology, two feature engineering methods are used including the bag of words and term frequency-inverse document frequency.

METHOD

Ensemble learning is widely used to attain high accuracy for classification tasks. The combination of various models can perform well as compared to individual models. Owing to the high accuracy of ensemble models, this study leverage an ensemble model to perform toxic comments classification. Our experiments indicate the good performance from Lstm and CNN, so to further improve the performance, this study combines these models. The proposed approach is called (Lstm-Cnn). it ensure that the class with a high predicted probability by two classifiers will be considered as the final prediction.

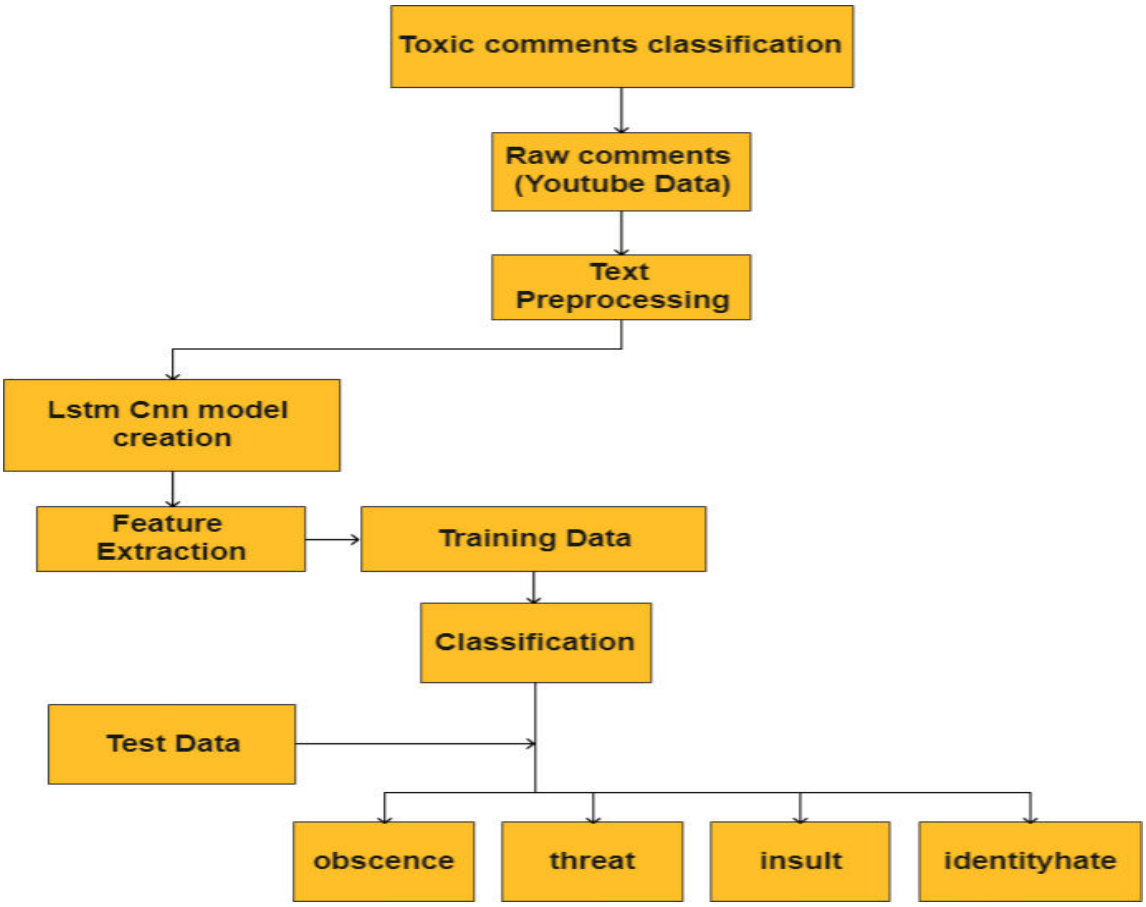

Fig 1: Architecture

IV. CONCLUSION

This study analyzes the performance of various machine learning models to perform toxic comments classification and proposes an

ensemble approached called Lstm-cnn. The influence of an imbalanced dataset and balanced datasetusing random under-sampling and over-sampling on the performance of the models is analyzed through extensive experiments. Two feature extraction approaches including TF-IDF a are used to get the feature vector for models' training. Results indicate that models perform poorly on the imbalanced dataset while the balanced dataset tends to increase the classification accuracy. Besides the machine learning classifiers like SVM, RF, GBM, and LR, the proposed RVVC and RNN deep learning models perform well with the balanced dataset. The performance with an over-sampled dataset is better than the under-sampled dataset as the feature set is large when the data is over-sampled which elevates the performance of the models. Results suggest that balancing the data reduces the chances of models over-fitting which happens if the imbalanced dataset is used for training. Moreover, TF-IDF shows better classification accuracy for toxic comments The proposed ensemble approach Lstm-cnn demonstrates its efficiency for toxic and non-toxic comments classificationThe performance of Lstm-cnn is superior both with the imbalanced and balanced dataset, yet, it achieves the highest accuracy of 0.97 when used with TF-IDF features The performance comparisonwith state-of-the-art approaches also indicates that Lstm-cnn shows better performance and proves good on small and large feature vectors. Despite the better performance of the proposed ensemble approach, its computational complexityis higher than the individual models which is an important topic for our future research. Similarly, dataset imbalance can overstate the results because data balancing using or random under-sampling approach may have a certain influence on the reported accuracy. Moreover, we intend to perform further experiments on multi-domain datasets and run experiments on more datasets for toxic comment classification.

V. REFERENCES

[1] E. Aboujaoude, M. W. Savage, V. Starcevic, and W. O. Salame, ‘‘Cyberbullying: Review of an old problem gone viral,’’ J. Adolescent Health, vol. 57, no. 1, pp. 10–18, Jul. 2015.

[2] How Much Data is Created on the Internet Each Day? Accessed:

Jun. 6, 2020. [Online]. Available: https://blog.microfocus.com/how- muchdata-is-created-on-the-internet-each-day/

[3] World Internet Users and 2020 Population Stats. Accessed: Jun. 6, 2020. [Online]. Available: https://www.internetworldstats.com/stats.htm

[4] M. Duggan, ‘‘Online harassment,’’ Pew Res. Center, Washington, DC, USA, Tech. Rep., 2014. [Online]. Available: https://www.pewresearch. org/internet/wp-content/uploads/sites/9/2017/07/PI

_2017.07.11_OnlineHarassment_FINAL.pdf

[5] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, ‘‘Deep learning for hate speech detection in tweets,’’ in Proc. 26th Int. Conf. World Wide Web Companion, 2017, pp. 759–760.

[6] Man Jailed for 35 years in Thailand for Insulting Monarchy on Facebook. Accessed: Jun. 6, 2020. [Online]. Available: https://www.theguardian. com/world/2017/jun/09/man-jailed-for-35-years-in-thailand-forinsulting- monarchy-on-facebook

[7] Mississippi Teacher Fired After Racist Facebook Post; Black Parent Responds. Accessed: Jun. 6, 2020. [Online]. Available: https://www. clarionledger.com/story/news/2017/09/20/mississippi-teacher-fired- afterracist-facebook-post/684264001/

[8] E. Wulczyn, N. Thain, and L. Dixon, ‘‘Ex machina: Personal attacks seen at scale,’’inProc.26thInt.Conf.WorldWideWeb,Apr.2017,pp. 1391–1399.

[9] M. Ptaszynski, J. K. K. Eronen, and F. Masui, ‘‘Learning deep on cyberbullying is always better than brute force,’’ in Proc. LaCATODA@ IJCAI, 2017, pp. 3–10.

[10] V. Srikanth. “ANALYZING THE TWEETS AND DETECT TRAFFIC FROM TWITTER ANALYSIS” v srikanth | INTERNATIONAL JOURNAL OF MERGING TECHNOLOGY AND ADVANCED RESEARCH IN COMPUTING, 20 MARCH. 2017. http://ijmtarc.in/Papers/Current%20Papers/IJMTARC-170309.pdf

[11] V. Srikanth. “A NOVEL METHOD FOR BUG DETECTION TECHNIQUES USING INSTANCE SELECTION AND FEATURE SELECTION” v srikanth | INTERNATIONAL JOURNAL OF INNOVATIVE ENGINEERING AND MANAGEMENT RESEARCH, 08 DECEMBER. 2017. https://www.ijiemr.org/public/uploads/paper/976_approvedpaper.pdf

[12] V. Srikanth. “SECURED RANKED KEYWORD SEARCH OVER ENCRYPTED DATA ON CLOUD” v srikanth | INTERNATIONAL JOURNAL OF INNOVATIVE ENGINEERING AND MANAGEMENT RESEARCH, 08 Febraury. 2018. http://www.ijiemr.org/downloads.php?vol=Volume-7&issue=ISSUE-02

[13] V. Srikanth. “WIRELESS SECURITY PROTOCOLS (WEP,WPA,WPA2 & WPA3)” v srikanth | Journal of Emerging Technologies and Innovative Research (JETIR), 08 mAY. 2019. https://www.jetir.org/papers/JETIRDA06001.pdf

[14] V. Srikanth, et al. “Detection of Fake Currency Using Machine Learning Models.” Deleted Journal, no. 41, Dec. 2023, pp. 31–38. https://doi.org/10.55529/ijrise.41.31.38.

[15] V. Srikanth, et al. “A REVIEW ON MODELING AND PREDICTING OF CYBER HACKING BREACHES.” 25 Mar. 2023, pp. 300–305. http://ijte.uk/archive/2023/A-REVIEW-

ON-MODELING-AND-PREDICTING-OF-CYBER-HACKING-BREACHES.pdf.

[16] V. Srikanth, "DETECTION OF PLAGIARISM USING ARTIFICIAL NEURAL NETWORKS." 25 Mar. 2023, pp. 201–209. http://ijte.uk/archive/2023/DETECTION-OF-PLAGIARISM-USING-ARTIFICIAL-NEURAL-NETWORKS.pdf.

[17] V. Srikanth, "CHRONIC KIDNEY DISEASE PREDICTION USING MACHINELEARNINGALGORITHMS." 25 January. 2023, pp. 106–122. http://ijte.uk/archive/2023/CHRONIC-KIDNEY-DISEASE-PREDICTION-USING-MACHINE-LEARNING-ALGORITHMS.pdf.

[18] Srikanth veldandi, et al. "View of Classification of SARS Cov-2 and Non-SARS Cov-2 Pneumonia Using CNN". journal.hmjournals.com/index.php/JPDMHD/article/view/3406/2798.

[19] Srikanth veldandi, et al. "Improving Product Marketing by Predicting Early Reviewers on E-Commerce Websites." Deleted Journal, no. 43, Apr. 2024, pp. 17–25. https://doi.org/10.55529/ijrise.43.17.25.

[20] Srikanth veldandi, et al."Intelligents Traffic Light Controller for Ambulance." Journal of Image Processing and Intelligent Remote Sensing, no. 34, July 2023, pp. 19–26. https://doi.org/10.55529/jipirs.34.19.26.

[21] Veldandi Srikanth, et al. "Identification of Plant Leaf Disease Using CNN and Image Processing." Journal of Image Processing and Intelligent Remote Sensing, June 2024, https://doi.org/10.55529/jipirs.44.1.10.

[22] S. Carta, A. Corriga, R. Mulas, D. Recupero, and R. Saia, "A supervised multi-class multi-label word embeddings approach for toxic comment classification," in Proc. KDIR, 2019, pp. 105–112.

[23] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in Proc. 1st Workshop Abusive Lang. Online, 2017, pp. 85–90.

[24] S. R. Basha, J. K. Rani, J. P. Yadav, and G. R. Kumar, "Impact of feature selection techniques in text classification: An experimental study," J. Mech. Continua Math. Sci., no. 3, pp. 39–51, 2019.

[25] S. R. Basha and J. K. Rani, "A comparative approach of dimensionality reduction techniques in text classification," Eng., Technol. Appl. Sci. Res., vol. 9, no. 6, pp. 4974–4979, Dec. 2019.

[26] M. S. Basha, S. K. Mouleeswaran, and K. R. Prasad, "Sampling-based visual assessment computing techniques for an efficient social data clustering," J. Supercomput., pp. 1–25, Jan. 2021.

[27] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, ‘‘Abusive language detection in online user content,’’ in Proc. 25th Int. Conf. World Wide Web. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, Apr. 2016, pp. 145– 153.

[28] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques, ‘‘Hate speech classification in social media using emotional analysis,’’ in Proc. 7th Brazilian Conf. Intell. Syst. (BRACIS), Oct. 2018, pp. 61–66.

[29] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, ‘‘Deceiving Google’s perspective API built for detecting toxic comments,’’ 2017, arXiv:1702.08138. [Online]. Available: http://arxiv.org/abs/1702.08138

[30] Toxic Comment Classification Challenge. Accessed: May 5, 2020. [Online]. Available: https://www.kaggle.com/c/jigsaw-toxic-commentclassification-challenge

[31] S. Alam and N. Yao, ‘‘The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis,’’ Comput. Math. Org. Theory, vol. 25, no. 3, pp. 319–335, Sep. 2019.

[32] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, ‘‘Tweets classification on the base of sentiments for US airline companies,’’ Entropy, vol. 21, no. 11, p. 1078, Nov. 2019.

[33] M. Anandarajan, C. Hill, and T. Nolan, Practical Text Analytics: Maximizing the Value of Text Data (Advances Analytics Data Science), vol.

2. Cham, Switzerland: Springer, 2019.

[34] Z. Z. Wint, T. Ducros, and M. Aritsugi, ‘‘Spell corrector to social media datasets in message filtering systems,’’ in Proc. 12th Int. Conf. Digit. Inf. Manage. (ICDIM), Sep. 2017, pp. 209–215.

[35] S. Yang and H. Zhang, ‘‘Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis,’’ Int. J. Comput. Inf. Eng, vol. 12, no. 7, pp. 525–529, 2018.

[36] F. F. Bocca and L. H. A. Rodrigues, ‘‘The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling,’’ Comput. Electron. Agricult., vol. 128, pp. 67– 76, Oct. 2016.

[37] J. Heaton, ‘‘An empirical analysis of feature engineering for predictive modeling,’’ in Proc. SoutheastCon, Mar. 2016, pp. 1–6.

[38] S. C. Eshan and M. S. Hasan, ‘‘An application of machine learning to detect abusive

Bengali text,'' in Proc. 20th Int. Conf. Comput. Inf. Technol. (ICCIT), Dec. 2017, pp. 1–6.

[39] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, and G. S. Choi, ''A performance comparison of supervised machine learning models for COVID-19 tweets sentiment analysis,'' PLoS ONE, vol. 16, no. 2, Feb. 2021, Art. no. e0245909.

[40] E. B. Fatima, B. Omar, E. M. Abdelmajid, F. Rustam, A. Mehmood, and

G. S. Choi, ''Minimizing the overlapping degree to improve classimbalanced learning under sparse feature selection: Application to fraud detection,'' IEEE Access, vol. 9, pp. 28101–28110, 2021.

Author Profiles

MS.G.SREELEKHA is currently working as Assistant Professor in Audisankara college of Engineering & Technology (AUTONOMOUS), NH-5, Bypass Road, Gudur, Tirupathi (Dt.), AndhraPradesh, India.

V.KOUSALYA is pursuing MCA from Audisankara college of Engineering & Technology(AUTONOMOUS), NH-5, Bypass Road, Gudur, Tirupathi (Dt.), AndhraPradesh , India.