Detection of Cyberbullying on Social Media using Machine Learning

Hanumanthu Venkata Kalyan Sampath[1], Mojjada Nandini[2], Joga Geetha Manjari[3], Boddu Padmasandhya[4]

[1,2,3,4] *Final Year B.Tech, CSE, Sanketika Vidya Parishad Engineering College, Visakhapatnam, A.P, India Guided by G.Sandhya, Associate Professor, SVPEC, Visakhapatnam, A.P, India*

Abstract - *With the rise of the Internet, the usage of social media has increased tremendously, and it has become the most influential networking platform in the twenty-first century. However, increasing social connectivity frequently causes problems. Negative societal effects that add to a handful of disastrous outcomes online harassment, cyberbullying, and other phenomena Online trolling and cybercrime Frequently, cyberbullying leads to severe mental and physical distress, especially in women and children, forcing them to try suicide on occasion. Because of its harmful impact, online abuse attracts attention. Impact on society Many occurrences have occurred recently all across the world. Internet harassment, such as sharing private messages, spreading rumors, etc., and Sexual comments As a result, the detection of bullying texts or messages on social media has grown in popularity. The data we used for our work were collected from the website kaggle.com, which contains a high percentage of bullying content. Electronic databases like Eric, ProQuest, and Google Scholar were used as the data sources. In this work, an approach to detect cyberbullying using machine learning techniques. We evaluated our model on two classifiers SVM and Neural Network, and we used TF-IDF and sentiment analysis algorithms for features extraction. This achieved 92.8% accuracy using Neural Network with 3-grams and 90.3% accuracy using SVM with 4- grams while using TF-IDF and sentiment analysis.*

Key Words: Cyberbullying, Hate speech, Personal attacks, Machine learning, Feature extraction, Sentimental analysis, Cybercrime, Neural Networks.

1. INTRODUCTION

SOCIAL Media is a collection of web-based programmes that let users to create and share user-generated content, built on the conceptual and technological basis of Web 2.0. People can gain access to a wealth of knowledge, as well as a quick way to communicate. However, social media can have bad consequences, such as cyberbullying, which can affect people's lives, particularly children and teenagers. Cyberbullying is described as aggressive, intentional activities taken against a victim through digital communication methods such as sending messages and making remarks. Cyberbullying on social media, unlike traditional bullying, which mainly occurs at school during face-to-face conversation, can occur anywhere at any time. Bullies have the freedom to damage their classmates' sentiments because they don't have to face them and can hide behind the Internet. Because we everyone, especially youngsters, are continuously connected to the Internet or social media, victims are easily exposed to harassment. The rate of cyberbullying victimisation varies between 10% and 40%. In the United States, almost 43% of teenagers have been bullied on social media at some point. It is our utmost priority to combat cyberbullying is to automatically detect and report bullying texts so that appropriate actions can be taken to avoid potential catastrophes.

Cyberbullies can be found in work or at school in the classic way. Bullying via cyberspace, on the other hand, stay anonymous, making this type of bullying both effective and harmful. Bullying in schools typically targets children who are physically weak, overweight, unpopular, or disabled, and the bullying occurs during the school day. There is no certain moment when a victim of cyberbullying will be bullied. As a result, the youngsters feel more victimised than usual. Bullying in cyberspace can take the form of uploading photographs or sending depreciate messages. and interactions that can take place in virtual reality, which differs from the reality we are used to. He or she may get a brief break from the bullying, but in cyberbullying, there is no relief from the tension until the victim returns the electronic device. The work of Dooley et al supports the victim's increased sense of powerlessness as a result of cyberbullying (2009). The same victim may predict when he or she will be bullied (for example, in school or on the playground), whereas a victim of cyber bullying has no idea when, when, or how he or she will be bullied (e.g. cell phone, computer), This causes an increased sense of powerlessness. According to recent studies, Online bullying is widespread and is among the most common forms of harassment among adolescents.

Cyberbullying is an arising societal issue in the digital period. The Cyberbullying Research Centre conducted a civil check of 5700 adolescents in the US and plant that 33.8 of the repliers had been cyberbullied and11.5 had cyberbullied others. While cyberbullying occurs in different online channels and platforms, social networking spots (SNSs) are rich grounds for online bullying. A recent check conducted by Ditch the Marker, an anti-bullying charity site, showed the frequency of cyberbullying on social networking spots (SNS bullying) 46 of the repliers reported being bullied, and 20 reported bullying others on SNSs. SNS bullying refers to any form of aggressive gestures on SNSs conducted by a group or an individual constantly and over time against targets who can not fluently defend themselves. It induces serious psychosocial and physical detriment similar to depression and tone-harming actions, with the most woeful outgrowth being self-murder. In one case, a teenage girl shot and killed herself after being relentlessly bullied on SNSs. Several crucial SNS features similar to digital biographies, relational ties, hunt, and sequestration, and network translucency give numerous openings for triple complementary relations between perpetrators, victims, and onlookers in SNS bullying incidents.

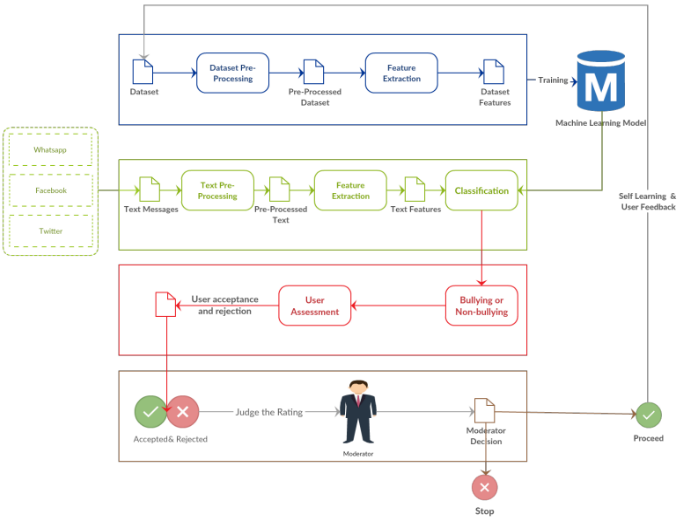
2. RELATED WORK

Another type of procedure uses Deep Learning and Neural Networks. One of the suggested methods is Zhang et al. in their paper uses a novel pronunciation-based convolution neural network (PCNN), thereby clearing the problem of noise and bullying data sparsity to overcome the class inequality. 1313 messages from Twitter, 13,000 messages from Formspring. me. The accurateness of the Twitter dataset wasn't computed due to it being imbalanced. While Achieving 56% precision, 78% recall, and 96% accuracy, while achieving high accuracy their dataset was unbalanced, which gives false results and that reflects in a precision score which is 56%. The authors Nobata et al. showed that using offensive language has increased recently, They used a framework called Vowpal wabbit for classification, and they also developed a supervised classification methodology with NLP qualities that outperform the deep learning approach, The F-Score reached 0.817 using dataset collected from comments posted on Yahoo News and Finance.

And Naïve Bayes 63. Moving on to Di Capua etal. they proposed a new way for cyberbullying discovery by espousing an unsupervised approach, they used the classifiers inconsistently over their dataset, applying SVM on FormSpring and achieving 67 on recall, using GHSOM on YouTube and achieving 60 perfection, 69 delicacy, and 94 recall, applying Naïve Bayes on Twitter and achieving 67 delicacy. Further, Haidar etal. proposed a model to descry cyberbullying but using the Arabic language they used Naïve Bayes and acquired 90.85 perfection and SVM achieved 94.1 perfection but they've a high rate of false-positive also the workshop on the Arabic language. Another type of system uses Deep Literacy and Neural Networks. One of the proposed styles is Zhang etal. in their paper uses a new pronunciation- grounded complication neural network (PCNN), thereby easing the difficulty of noise and bullying data sparsity to overcome the class imbalance. 1313 dispatches from Twitter, dispatches from Formspring. me. The delicacy of the Twitter dataset wasn't reckoned due to it being imbalanced. While Achieving 56 perfection, 78 recall, and 96 rigorousness, while achieving high delicacy their dataset was unstable, which gives false results and that reflects in a perfection score which is 56. The authors Nobata etal. showed that using vituperative language has boosted lately, They used a frame called Vowpal wabbit for bracket, and they also developed a supervised bracket methodology with NLP features that outperform the deep literacy fashion, The F- Score reached 0.817 using dataset collected from commentary posted on Yahoo News and Finance. Zhao etal. proposed a frame-specific for cyberbullying discovery, they used word embedding that makes a list ofpre-defined insulting words and assign weights to gain bullying features, they used SVM as their main classifier and got a recall of 79.4. Also another approach was proposed by Prime etal. they got their dataset from MySpace and manually marked them and they used the SVM Classifier for the bracket. Also, Chen etal. proposed a new point birth system called Lexical Syntactic Point and SVM as their classifier and they achieved 77.9 perfection and 77.8 recall. Likewise, Ting etal. proposed a strategy grounded on SNM, they collected their data from social media and also used SNA measures and sentiments as features. Seven trials were made and they achieved around 97 perfection and 71 recall. Likewise, Harsh Dani etal. introduced a new frame called SICD, they used KNN for bracket. Eventually, they achieved a 0.6105 F1 score and 0.7539 AUC score. The SVM classifier was one of the approaches used in the exploration papers. Dadvar etal. have constructed in the first and alternate papers a Support Vector Machine classifier using WEKA, their dataset was collected from Myspace. They achieved 43 on perfection, and 16 in recall and they did n't mention the delicacy, the only difference between the two papers is that they used gender information in order in the alternate paper. Also, in their alternate paper 4626 commentary from 3858 distinct druggies were collected. The commentary were manually labeled as bullying (9.7) andnon-bullying (inter-annotator agreement 93). The SVM classifier was used by them and was suitable to reach results of over to 78 on perfection and 55 on recall. Eventually, in their third paper, they used 3 models for their dataset gathered from the YouTube comment section Multi-Criteria Evaluation Systems (MCES), machine literacy (Naïve Bayes classifier, decision tree, SVM), and Mongrel approach. The MCES score 72 on the delicacy, while Naïve Bayes scored the loftiest out of the three with 66. Moving on to another author, Potha etal. have also used the SVM approach and fulfilled a 49.8 result on rigorousness. While Chavan etal. used two classifiers logistic retrogression and support vector machine. The logistic retrogression achieved 73.76 rigor and 60 recall and64.4 Precision. While for the support vector machine they achieved 77.65 delicacy and 58 recall and 70 perfection and they got their dataset from Kaggle:

3. PROPOSED APPROACH

The proposed technique, as visible in Fig. 1, consists of 3 principal steps: Preprocessing, capabilities extraction, and category step. In the preprocessing step we smooth the information through deducting the noise and disproportionate textual content. The preprocessing step is achieved within side the following: - Tokenization: In this part, we take the textual content as sentences or complete paragraphs after which output the entered textual content as separated phrases in a listing. - Lowering textual content: This takes the listing of phrases that were given out of the tokenization after which lowers all of the letters Like: 'THIS IS AWESOME' goes to be 'that is awesome'. - Stop phrases and encoding cleaning: This is a essential a part of the preprocessing wherein we smooth the textual content from the ones forestall phrases and encode characters like n or t which do now no longer offer a bit of significant data to the classifiers. - Word Correction: In this part, we used Microsoft Bing phrase correction API that takes a phrase after which returns a JSON item with the maximum comparable phrases and the space among those phrases and the unique phrase



The 2d step of the proposed Model is the capabilities extraction step. In this step, the textual information is transformed right into a appropriate layout relevant to feed into gadget getting to know algorithms. First, we extract the capabilities of the enter information the usage of TFIDF and placed them in a capabilities listing. The key concept of TFIDF is that it really works at the textual content and receives the weights of the phrases regarding the record or sentence. In Addition to TFIDF, we use the sentiment evaluation technique to extract the polarity of the sentences and upload them as a function to the capabilities listing owning the TFIDF capabilities. The polarity of the sentences method that if the sentence is assessed as high-quality or negative. For that purpose, we extract the polarity the usage of the Text Blob library That's a pre-educated version for film reviews. In more to the function extraction the usage of TFIDF and sentiment polarity extraction, the proposed method makes use of NGram[28] to keep in mind the one of a kind assortments of the phrases for the duration of the assessment of the version. Particularly, we use used 2- Gram, 3-Gram, and 4-Gram. The remaining step withinside the proposed technique is the category step wherein the extracted capabilities are fed right into a category set of rules to train, and check the classifier and consequently use it withinside the prediction phase. We used classifiers, namely, SVM (Support Vector Machine) and Neural Network. The neural community consists of 3 layers: Input, hidden, and output layers. In the enter layer, it includes 128 nodes. In the hidden layer, it consists of sixty four neurons. The output layer is a Boolean output. Generally, the assessment of classifiers is achieved the usage of numerous assessment matrices relying at the confusion matrix. Among the ones standards are Accuracy, precision, recall, and f-score. They are calculated consistent with the subsequent equations:

$$\text{Accuracy} = \frac{T\,P + T\,N}{T\,P + T\,N + F\,P + F\,N} \quad (1)$$

$$\text{Precision} = \frac{T\,P}{T\,P + F\,P}$$

$$\text{Recall} = \frac{T\,P}{T\,P + F\,N}$$

$$F - \text{Score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Where TP represents the number of true positive, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives classes.

4. EXPERIMENTAL RESULTS

This section describes the experimental results on the proposed approach. We evaluate the proposed approach on the cyberbullying dataset from kaggle. within the following we describes the information and also the results.

A. Data Description

We have used cyberbullying dataset from Kaggle which was collected and labeled by the authors Kelly Reynolds et al. in their paper [2]. This dataset contains normally 12773 conversations messages collected from Formspring. The dataset contains questions and their answers annotated with either cyberbullying or not. The annotation classes were unbalanced distributed such 1038 question-answering instances out of 12773 belongs to the category cyberbullying, while 11735 belongs to the opposite class. First, to remedy the information unbalancing, we take the identical number instances of both classes to live the accuracy. We also faraway from the information big size conversations and remove the noisy data. We ended up with total 1608 instance conversations where 804 instances belongs to every class. Table I summarizes the statistics of dataset.

TABLE I. STATISTICS OF THE DATASET

Total number of Conversations	1608
Number of cyberbullying	804
Number of non-Cyberbullying	804
Number of distinct words	5628
Number of token	48843
Maximum Conversation size	773 Characters
Minimum Conversation size	59 Characters

B. Results

After preprocessing the dataset, we follow the identical step presented in Section III to extract the features. We then split the dataset into ratios (0.8,0.2) for train and test. Accuracy, recall and precision, and f-score are taken as a performance measure to judge the classifiers. We apply SVM likewise as Neural Network (NN) as they're among the most effective performance classifiers within the literature. We run several experiments on different n-gram language model. particularly, we take into consideration 2-gram, 3-gram, and 4-gram during the evaluation of the model produced by the classifiers. Table II summarizes the accuracy of both SVM and NN. The SVM classifier achieved the best percentage using 4-Gram with accuracy 90.3% while the NN achieved highest accuracy using 3-Gram with accuracy 92.8%. it's found that the common accuracy of all n-gram models of NN achieves 91.76%, while the average accuracy of all n-gram models of SVM achieves 89.87%. Fig. 2 depicts the accuracy results of both classifiers

TABLE II. THE ACCURACY OF SVM AND NN IN DIFFERENT LANGUAGE MODEL

Classifier	2-Gram	3-Gram	4-Gram	Average
SVM	89.42%	89.9%	**90.3%**	89.87%
Neural Network	90.9%	**92.8%**	91.6%	91.76%

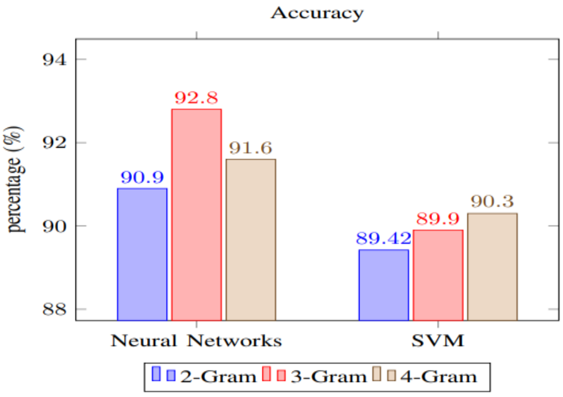


Fig. 2. Comparison between SVM and Neural Network in Terms of Accuracy

n addition to accuracy, Table III and Table IV show the evaluations of both classifiers in terms of precision and recall respectively for every language model. The trade-off between recall and precision is shown in Table V which represents the fscore of both classifiers within the different language model. Table V summarizes the f-score of both SVM and NN. The SVM classifier achieved the very best f-measure using 4-Gram with f-score 90.3% while the NN achieved highest f-measure using 2-Gram with f-score 92.2%. it's found that the common f-score of all n-gram models of NN achieves 91.9%, while the typical f-score of all n-gram models of SVM achieves 89.8%. Fig. 3 summarizes the f-score of the classification of the SVM and Neural Network. The results of average accuracy furthermore because the average f-score indicate that NN performs better than SVM.

TABLE III. RECALL OF SVM AND NN

Classifier	2-Gram	3-Gram	4-Gram	Average
SVM	89.42%	90.3%	**90.8%**	90.1%
Neural Network	91.6%	91.5%	**92%**	91.7%

TABLE IV. PRECISION OF SVM AND NN

Classifier	2-Gram	3-Gram	4-Gram	Average
SVM	89.42%	89.5%	**90%**	89.6%
Neural Network	**93%**	92.5%	91.7%	92.4%

TABLE V. F-SCORE OF SVM AND NN

Classifier	2-Gram	3-Gram	4-Gram	Average
SVM	89.42%	89.8%	**90.3%**	89.8%
Neural Network	**92.2%**	91.9%	91.8%	91.9%

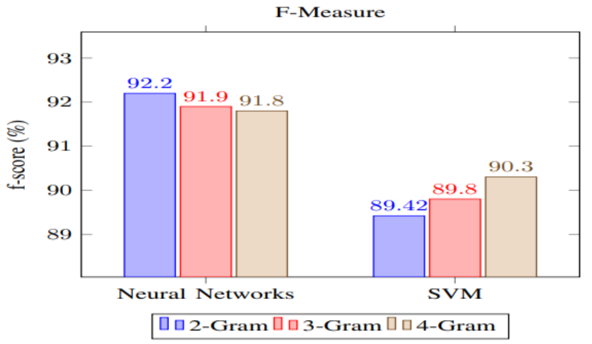


Fig. 3. Comparison between SVM and Neural Network in Terms of F-Measure

Additionally to the previous experiments, we evaluate and compare our classifiers on the proposed approach with the work. during this work, they used logistic regression and SVM for classification and used the identical data. Moreover, we have calculated the typical accuracy, recall, precision and Fscore of our two classifiers. The summary of results is shown in Table VI. to check the work, it's found that our proposed NN model outperforms all other classifiers and is ranked as the best leads to terms of average accuracy and F-Score achieving accuracy 91.76% and f-score 91.9%. In Fig. 4 we are comparing between our greatest classifier with their best classifier in case of accuracy. Finally, here in Fig. 5 we are comparing between our greatest classifier with their best classifier just in case of F-Measure.

TABLE VI. COMPARISON WITH RELATED WORK

	Classifier	Avg. Accuracy	Avg. Recall	Avg. Precision	Avg. F-Score
Vikas S Chavan	Logistic regression	73.76	61.47%	64.4%	62.9%
	SVM	77.65%	58.29%	70.29%	63.7%
Current Results	Neural Network	**91.76%**	**91.7%**	**92.4%**	**91.9%**
	SVM	89.87%	90.1%	89.6%	89.8%

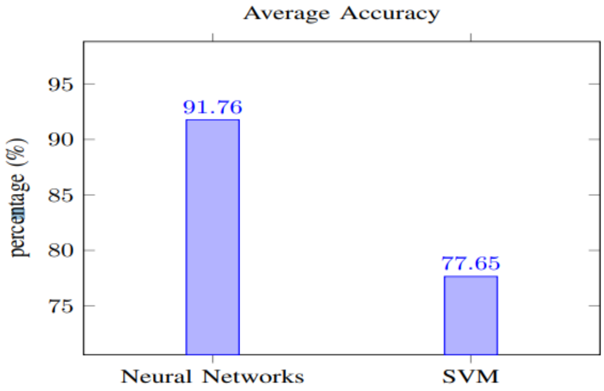


Fig. 4. Comparison between the Best Classifiers in Terms of Accuracy

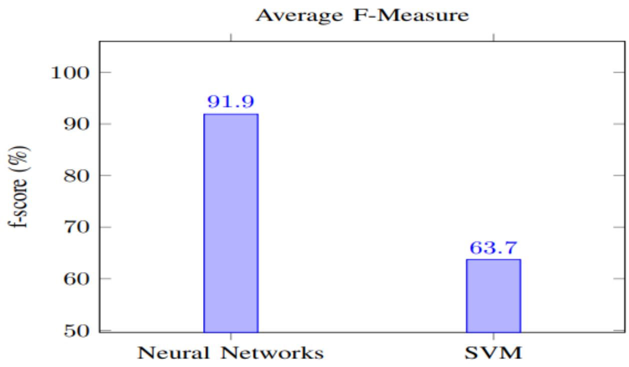


Fig. 5. Comparison between the Best Classifiers in Terms of F-Measure

5. CONCLUSION

In this paper, we proposed an approach to descry cyberbullying using machine literacy ways. We estimated our model on two classifiers SVM and Neural Network and we used TF IDF and sentiment analysis algorithms for features birth. The groups were estimated on different n-gram language models. We achieved 92.8 delicacy using Neural Network with 3-grams and 90.3 delicacy using SVM with 4-grams while using both TF IDF and sentiment analysis together. Here, our Neural Network performed better than the SVM classifier as it also achieves an average f- score of 91.9 while the SVM achieves an average f- score of 89.8. Likewise, we compared our work with another affiliated work that used the same dataset, chancing that our Neural Network outperformed their classifiers in terms of delicacy and f- score. By achieving this delicacy, our work is going to ameliorate cyberbullying discovery to help people to use social media safely.

6. FUTURE SCOPE

By obtaining this level of precision, cyberbullying detection will undoubtedly increase, allowing people to utilize social media safely. The amount of the training data, however, limits the detection of cyberbullying patterns. To increase the performance, more cyberbullying data is required. As a result, deep learning techniques will be appropriate for larger data because they have been shown to outperform machine learning algorithms on larger datasets. Cyberbullying can be accurately detected using data sets acquired from multiple sources. The scarcity of data should be reduced.

7. REFERENCES

[1] C. Fuchs, Social media: A critical introduction. Sage, 2017.

[2] N. Selwyn, “Social media in higher education,” The Europa world of learning, vol. 1, no. 3, pp. 1–10, 2012.

[3] H. Karjaluoto, P. Ulkuniemi, H. Keinanen, and O. Kuivalainen, "An- ¨ tecedents of social media b2b use in industrial marketing context: customers' view," Journal of Business & Industrial Marketing, 2015.

[4] W. Akram and R. Kumar, "A study on positive and negative effects of social media on society," International Journal of Computer Sciences and Engineering, vol. 5, no. 10, pp. 351–354, 2017.

[5] D. Tapscott et al., The digital economy. McGraw-Hill Education,, 2015.

[6] S. Bastiaensens, H. Vandebosch, K. Poels, K. Van Cleemput, A. Desmet, and I. De Bourdeaudhuij, "Cyberbullying on social network sites. an experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully," Computers in Human Behavior, vol. 31, pp. 259–271, 2014.

[7] D. L. Hoff and S. N. Mitchell, "Cyberbullying: Causes, effects, and remedies," Journal of Educational Administration, 2009.

[8] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," Archives of suicide research, vol. 14, no. 3, pp. 206–221, 2010.

[9] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7, 2009.

[10] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in In Proceedings of the Social Mobile Web. Citeseer, 2011.

[11] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in 2011 10th International Conference on Machine learning and applications and workshops, vol. 2. IEEE, 2011, pp. 241–244.

[12] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using twitter users' psychological features and machine learning," Computers & Security, vol. 90, p. 101710, 2020.

8. BIOGRAPHIES

G.SANDHYA

Currently working as assistant professor from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering college

HANUMANTHU VENKATA KALYAN SAMPATH

Pursuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College

MOJJADA NANDINI

Pursuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College

JOGA GEETHA MANJARI

Pursuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College

BODDU PADMA SANDHYA

Pursuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College