AN ENSEMBLE LEARNING APPROACH FOR IDENTIFYING A FAKE NEWS DETECTION

Mr.V Chandrasekhar[1] , A.V.Pavan Kumar[2]
[1]Assistant Professor, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS), Gudur, AP, India.
[2]PG Scholar, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS), Gudur, AP, India.

Abstract - Science and computer scientists find fake news identification to be a fascinating subject. More people than ever before are generating and sharing information online. Users all across the world are exposed to a large number of false articles from various sources. The majority of these stories are inaccurate, deceptive, or unrelated to reality. Automatically identifying text as rumour or fake is difficult. It is also difficult to analyse the text and classify news headlines or fake articles for reporting reasons. Several variables affect the text's veracity. The task will involve classifying news articles automatically using an ensemble technique. To condense the information in lengthy articles and generate knowledge, textual attributes will be detected using the NLTK toolkit. The created ensemble methods-based technique will be tested using real-world data and will offer a high level of accuracy as well as contribute to improving the individual models' capacity for prediction.

Key Words: Ensemble, Fake news, Detection, Machine Learning

1. INTRODUCTION

Since many years ago, there has been an increase in the problem of false information, especially with the use of social media growing so rapidly[10]. False news is frequently disseminated to influence certain organisations or politicians. Furthermore, fake news is frequently shared on the internet by individuals who seek to harm a person, organisation, or agency in order to benefit politically or financially[10]. Furthermore, false rumours circulated to give individuals false impressions of the truth[3]. The rise of false information or fake news is surely not another wonder. It turns out to be more understandable at times when there is more media coverage, such as during the 2016 demonetization initiative in India[11]. Exploration has ordinarily uncovered that out of all online media stages, Twitter does well in uncovering improper information as a result of oneself altering properties of publicly supporting as clients share notions, theories, and proof[11]. It has been observed that, nowadays, numerous types of fake news are currently present on different types of social media objectives and it is a very hard task to tackle them[3]. There are countless systems which can detect whether the news is real or fake, are available but every system has content related issues. The first priority to classify the dataset is the efficient detection systems. Inthis paper, we assess the exhibition of ensemble approach for counterfeit news recognition on two datasets, one containing customary online news articles and the second one, news from different sources.

LITERATURE REVIEW

There has long been a market for false news reports. According to reports, the prevalence of false news impacted the outcome of the 2016 US Presidential Elections. Stefan and Paulheim [4] used this electioncoherence ratings as the foundation for their fake news identification. They also took into account the statistical aspects of the photographs, such as count and image ratio. They were accurate to a degree of roughly 83 percent.

2. METHODOLOGY

2.1 PREPROCESSING

2.1.2 Normalization: It is a method of deleting or replacing digits, punctuation, unused white spaces, and default stop words from text in order to reduce the amount of our data's vocabulary[6]. These elements have no bearing on how a statement is understood. The NLTK library has a number of stop words for each language that can be used to remove them from text and produce a list of word tokens[1].

2.1.2 Tokenization: Tokenization is the process of dividing an input sequence into discrete, meaningful pieces. These tokens are practical building blocks for additional semantic processing[1]. It might be a single word, phrase, paragraph, etc. The NLTK libraries contain a variety of tokenizers,including WhiteSpaceTokenizer, WordPunctTokenizer, TreebankWordTokenizer, etc. While TreebankTokenizer employs a different set of grammatical rules to tokenize inputs[6], WhiteSpaceTokenizer and WordPunctTokenizer divide the input using white spaces and punctuation, respectively.

Input = [“ensemble based approach for detection of fake news using machine learning”]

Output=[“ensemble”, ”based”, ”approach”, ”for”, ”detection”, ”of”, ”fake”, ”news”, ”using”, ”machine”, ”learning”]

2.1.3 Stemming: The Stem[1] refers to the process of obtaining a word's basic form after suffixes have been added or subtracted. It gathers several words into one block (stem). The NLTK library has a number of stemmers, including PorterStemmer, LancasterStemmer, and others. A less aggressive algorithm called PorterStemmer uses five rules for various scenarios that are gradually used to develop fundamental knowledge[1,6]. Although the stems it produces are frequently not true English words, they are intelligible. Additionally, it applies rules that are based on algorithms to generate stems rather of maintaining a lookup table. Iterative behaviour is a characteristic of the LancasterStemmer algorithm type[6]. In LancasterStemmer, rules are externally preserved. Due to the possibility of over-stemming from more repetitions, stems may become meaningless or cease to be linguistic. In addition, it could make things harder to understand.

2.1.4 Lemmatization: We often make things right by employing vocabulary and morphological analysis[1,6]. This procedure, which is known as a lemma, returns a word's base or dictionary form. The WordNet Lemmatizer tool from NLTK allows users to look for lemmas for certain terms by accessing a WordNet database. A lemmatizer, for instance, maps the verbs gone, going, and went to its standard form go.

2.2 ENSEMBLE LEARNERS

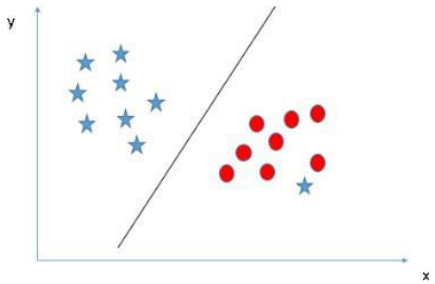
2.2.1 Naïve Bayes: It is one of the machine learning algorithms used for text classification problems[12]. Apart from that, it is very easy to implement and very efficient at the same time. There are three event models:

- Multivariate Bernoulli Event Model
- Multivariate Event Model
- Gaussian Naïve Bayes classification

Additionally, Nave Bays denotes that all features are independent of one another and that the presence of one feature has no bearing on the likelihood that another feature will appear. When given tasks with a small dataset, this model performs better than any other strong model[12].

A feature vector in multinomial naive bayes contains a term that represents the frequency with which the provided term occurs. Contrarily, Bernoulli is a binary classification that indicates if a term is there or not, and Gaussian is a continuous classification..

2.2.2 SVM (Support Vector Machine): Regression is classified using SVM, one of the supervised machine learning models[7]. However, categorization issues make extensive use of it. In this approach, the value of every single piece of data is often represented as a point on an N-dimensional space, where N is the total number of spaces we have[7]. The value of each element, which stands for the value of a particular coordinate, is also carried[13]. We eventually locate the hyper-plane to categorise the values. The SVM algorithm, for instance, can ignore outliers when locating the hyper-plane[12].



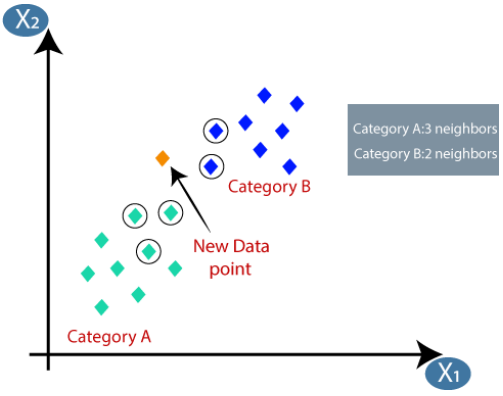
The SVM algorithm has numerous clear benefits, including:

- When it comes to categorising classes with a distinct and wide margin, it works quite well.
- Very efficient in big areas[12]. Additionally, if there are more dimensions than samples, the method is efficient.
- SVM uses less memory.

2.2.3Logistic Regression: It is a very simple machine learning classification model. The primary application of it is to forecast binary output, such as 0 or 1, Yes or No[7,12]. The independent variable is inputted. In cases when the predicted variable is a binary variable, it is also regarded as a particular case of the linear regression model. In plain English, this model predicts the likelihood that an event will occur. The probability is consistently between 0 and 1.

- Given that it is a relatively simple model, it has only the following advantages:
- It's incredibly simple to put into practise.
- 2 It strongly contradicts the premise.
- 3 It can be applied to several classes.
- Very fast when it comes to working with the unknown records compared to others.

2.2.3 KNN (K-Nearest Neighbors): It is one of the few ML models that is both supervised learning-compliant and among the simplest. It anticipates the similarities between the data for which we are forecasting the class and the existing classes, and at the conclusion[7], it places the new case into the category of the class that is highly similar to the record or data. It is applicable to both classification and regression. But classification[12] is where it is most frequently utilised. In addition, it is a lazy learner because it doesn't memorise anything or use training data to do so. Instead, it begins working as soon as it receives data that it needs to predict, which prevents it from utilising any memory.. It is also good atoutliers for example:



Given that we are employing KNN, the new data point may be regarded as an outlier in this case and is easily categorised into class A. Talking about the advantages:

- First of all, it is quite simple to put into practise.
- Data can be updated whenever you want because it is a slow learner.
- It saves time because there is no training period required..

Random Forest algorithm: This supervised machine learning algorithm is well-liked. It can be applied to both

classification and regression problems[7]. It is obviously based on the ensemble learning approach, which combines the output from various classifiers to increase overall accuracy. The Random Forest algorithm's most crucial characteristics are: It runs faster than other algorithms.

Even with a large dataset, it will always provide the highest accuracy of any model.

Even though the data set provides excellent accuracy, it is likely that there are occasionally missing data.

We can therefore conclude from the foregoing claims that it is quite competent of handling the high dimensional dataset and working for both regression and classification.However, it is more suitable for the regression problems.

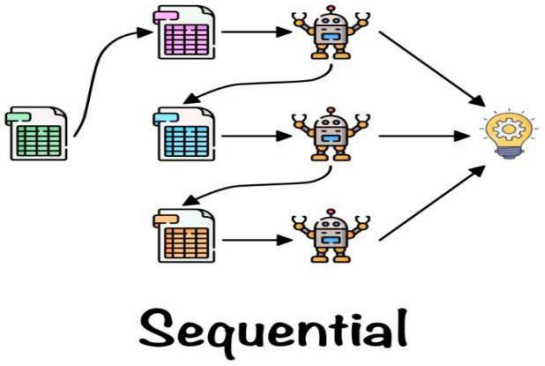
2.3 ENSEMBLE APPROACH

Ensemble method uses all of them to boost the first level models' ability to predict. Additionally, methods for ensemble learning are referred to as second level models.

Some models are effective at displaying the knowledge, while others are effective at demonstrating others[2]. Find several different individual models, then combine the outcomes to find the ideal one. The general robustness of each model makes up for its unique biases and aberrations. This offers a composite prediction whose accuracy is higher than the accuracy of the constituent models. There are two Ensemble methods:

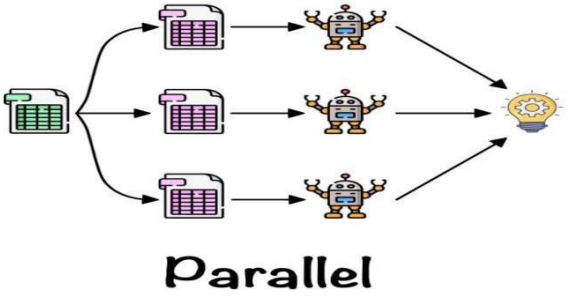
2.3.1 Sequential ensemble methods:

- Individual models appear one after the other.
- To benefit from ensemble learners (Individual Models) is the major goal[2].
- The overall performance of a group of students may surge.



2.3.2 Parallel ensemble methods:

- All individual models run simultaneously in parallel.
- The final outcome takes into account the class with the highest average probability.



The use of numerous models in an ensemble allows for performance gains beyond those possible with a single model.

- Robustness: Ensemble models take into account all of the base learners' predictions.
- Accuracy: Ensemble models perform better and make accurate predictions[2].

Depending on the dataset, each distinct model offers benefits and drawbacks. An ensemble technique is the ideal way to integrate all the benefits of each individual model to achieve the desired accuracy and boost predictive power since no single model can discover all the patterns from the dataset.

Ensemble Learning has multiple methods from which someare mentioned below:

Voting: It uses a parallel ensemble method for voting.

Step 1: Offer unique training data.

Step 2:Build and fit various classifiers to each of these varied copies in step two[2].

Step 3: Create as many forecasts as you can in order to create the final overall forecast.

We provide ensemble learners with a thorough dataset for voting (Individual models). Voting models are distinct for each ensemble learner. Since it is a parallel approach, each ensemble learner submits their output simultaneously, and the ensemble selects the majority-approved final output (individual o/p=1 1 1 0, Final o/p=1). Hard voting and soft voting are the two different ways of voting.

- In terms of voting, "hard voting" is what I described above, but "soft voting" has a taste akin to "bagging" and "boosting."
- Soft voting is based on summing the anticipated probabilities of classes and forecasting the class with the highest sum probabilities, whereas hard voting is entirely based on majority.

Bagging: It is an ensemble method used in parallel. By taking the mean of several estimates, bagging or bootstrap aggregation[2] decreases the variance of an estimate.

Create randomly picked datasets from the initial training data in step 1. (bootstrapping).

Step1:-Build a classifier and fit it to each of these various copies in step two.tep2

Step2:-Take the average of all forecasts in Step 3 to arrive at the final, comprehensive forecast.

Step3:-In bagging, we provide ensemble learners' (individual models') subset datasets, which are randomly selected data. The same bagging model is being used by all ensemble learners. Due to the parallel nature of the approach, all ensemble learners contribute their work at once, and the ensemble selects the final product with the highest probabilities (soft voting).

ii. Boosting: It is a sequential ensemble method and reduces bias by training weak learners sequentially, each trying to correct its predecessor. Boosting is a method for transforming frail learners into solid ones[3]. Each new tree is based on a slightly altered version of the first dataset.

- Step 1:- Train a classifier H1 that best classifies the data with respect to accuracy.
- Step 2:- Identify the regions where H1 produces errors, add weights to them, and produce a H2 classifier.
- Step 3:- Aggregate those samples for which H1 gives a different result from H2 and produces H3 classifier.

Boosting has two methods listed below:

- AdaBoost is the first boosting algorithm that has been used to problem-solving techniques. Weight records that were misclassified are rectified. Weight for inaccurate forecasts rises if models make poor predictions, and vice versa. With the help of changing weight, learning takes place. Two leaves typically cover its tree.
- Gradient boosting is a method for training several models successively, additively, and over time. By including weak learners, it employs the gradient descent technique to lower the loss function (MSE) of a model. With algorithms, Create a fundamental model, a typical ensemble learner, or the class that is utilised the most. Based on the projected average value and the actual value, analyse the residual error. We now develop a further RMI model with residuals as a target[3]. A fresh prediction residual has emerged. We will now determine the updated target anticipated value. The new RM 2 model will then match the target residuals a second time and forecast the new residuals now that we know the residuals (actually predicted).

3. PROPOSED SYSTEM

1. Dataset Description: On the basis of a data set containing headlines from several news websites, the suggested strategy was assessed. This dataset's main objective was to classify the proportion of fake and real rumors[3]. The dataset is balanced since it includes a variety of occurrences, including news about immigration, politics, social issues, healthcare, and more. The features of the news dataset are shown in the diagram below. There are about 23000 occurrences overall. Attributes of dataset -

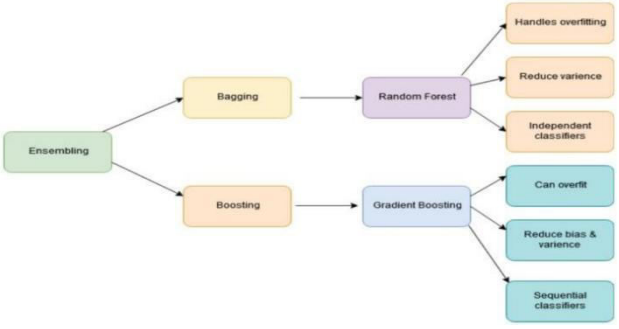
- News_id – the index of individual news.
- News_url – source of distributor of news.
- Title – The short title text should arouse the reader's interest and introduce the main theme of this article.
- Result – label of particular news(real or fake)
- **Feature Selection:**
- Since all ensemble learners only work with integers, we have created the algorithms listed below to transform strings to integers.
- TF-IDF:Term-frequency time (TF-IDF) Document-frequency inversion. We can infer that it establishes a formula for determining how important a word is to a text document[8]. It only adds special values to characteristics and informs us of the rarity of the word[8].

$$tf - idf(t,d) = tf(t,d) * \left(\frac{n}{df(t)} + 1\right)$$

- Count-Vectorizer: Each text sample from the document is represented as a row in the matrix by the count-vectorizer, which creates a grid where each intriguing word is addressed by a column of the words (vertically). Each cell's value is just the frequency of a specific word in the text document[3].

2. Parameter Selection:

- Multinomial naïve-bayes(alpha)
- Logistic Regression (C=100)
- Linear SVM (C=0.25)
- KNN(neighbor=120)
- Random Forest Classifier (number of feature=4)



3. RESULTS

Dry run was performed on several classifiers and their corresponding results were noted[5]. Our observations are appropriately illustrated in this article. Figure 1 display the The classification performance of experiments with different machine learning models (Multinomial Naive Bayes, Logistic Regression, Support Vector Machine, K - Nearest Neighbors, Random Forest)[2,3]. In our evaluation, we tracked down that all machine learning classifiers

accomplished approximately 85% and above accuracy.

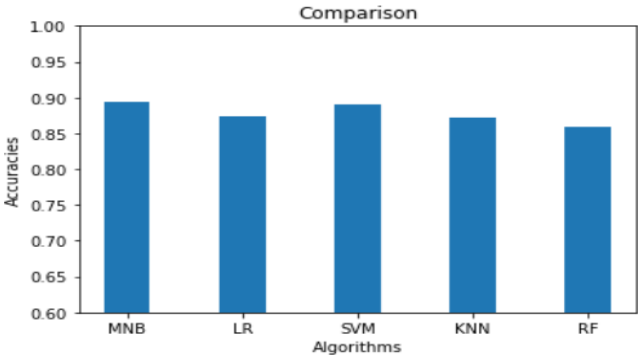


Fig.1. Accuracy

We had implemented svm bagging, Gradientboosting and Hard Voting and got accuracy 89%, 87%, 88% respectively. According to results, bagging with svm took crown place with maximum accuracy among three which means parallel ensemble approach is more beneficial for our dataset. We also measured the confidence score for all and it varies between 0 to 1 because it shows probability. If data with a high confidence score it means that it has high probability to take a place as a final result. Exactness is characterized as the exhibition of our learning calculation for right expectations[5].

Figure 2 and 3 shows the performance of the models. Apart from that it also indicates that the SVM Bagging and the Hard Voting gave output very clear and very effective also, with the TF-IDF[3]. From the figure below, we can get to know the training time of the individual algorithms varies from model to model. We can observe that training time is inversely proportional to number of iterations.

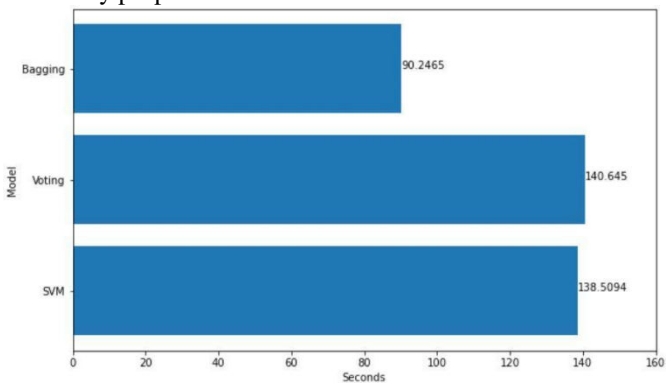


Fig.2. Time taken models

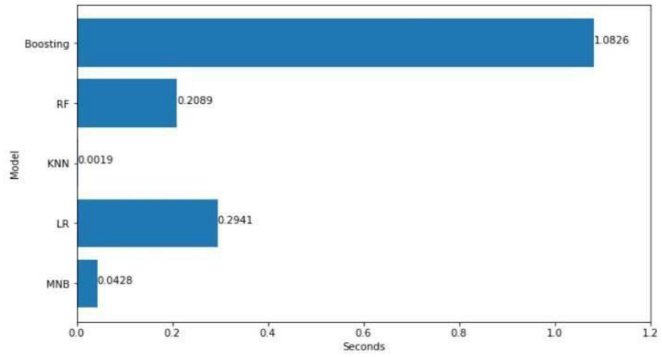


Fig.3. Time taken models

4. CONCLUSION

In our study, we're attempting to develop a machine learning model that will accurately differentiate authentic communications from false messages. It provides a general methodology and several elements upon which the veracity of the information is predicated. In our project, we used monitoring and deep learning techniques. The task of categorising news requires top-to-bottom knowledge in order to identify anomalies in the material. In this study, we discussed the complexity of fake headlines using ensemble approaches. To receive the most recent news, the information we utilise in our work is gathered from the Internet and includes headlines from a variety of industries. Finding patterns in text data that distinguish bogus articles from authentic news is a crucial component of this research[3,9]. Bagging which is a method of ensemble learning got 89% accuracy to recognize whether news is fake or real, which is 4% higher than the Random Forest Algorithm which is the worst performance among all individuals.Moreover, it has been extracted from research that ensemble approach performs well rather than individual models.

ACKNOWLEDGEMENT

We are grateful for this rare opportunity to express our appreciation to each and every one of the people who played a crucial role in the successful completion of our project.

We have discovered this uncommon chance to manifest an expression of gratitude to every one of the individuals who assumed a key job in the fruitful consummation of our undertaking.

We offer profound thanks to mentor Prof. Hemant Yadav inner task manage from Faculty of Engineering, CHARUSAT for their significant recommendations, help and good help. We also want to thank everyone who, despite not being able to think of another name, still genuinely and indirectly assisted.

We are really grateful to Dr. Abdul Jhummarwala, who served as our external guide and is a Research Scientist at BISAG-N in Gandhinagar. His wide subject knowledge and encouraging demeanour gave us the motivation to succeed. He was truly our protector during this time, providing the ideal balance of strict discipline and unwavering concern.

REFERENCES

[1] C. V. Gonzalez Zelaya, "Towards Explaining the Effects of Data Preprocessing on Machine Learning," 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019, pp. 2086-2090, doi: 10.1109/ICDE.2019.00245.

[2] Arush Agarwal, Akhil Dixit. “Fake News Detection: An Ensemble Learning Approach”. Published in 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 19 June 2020
[3] Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang. “Multiclass Fake News Detection using Ensemble Machine Learning”. Published in 2019 IEEE 9th International Conference on Advanced Computing (IACC). Tiruchirappalli, India. 30 January 2020.
[4] Weijie Jiang, Xingyou Wang, Zhiyong Luo.”Combination of convolutional and recurrent neural network for sentiment analysis of short texts.” COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers 2016.
[5] Ghosal, D., Bhatnagar, S., Akhtar, M.S., Ekbal, A. and Bhattacharyya, P., 2017. IITP at SemEval-2017 task 5: an ensemble of deep learning and feature based models for financial sentiment analysis. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval2017) (pp. 899-903)
[6] D. H. Deshmukh, T. Ghorpade and P. Padiya, "Improving classification using preprocessing and machine learning algorithms on NSL-KDD dataset," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), 2015, pp. 1-6, doi: 10.1109/ICCICT.2015.7045674.
[7] Ghosh, Souvick, and Chirag Shah. ”Towards automatic fake news classification.” Proceedings of the Association for Information Science and Technology 55, no. 1 (2018): 805-80.
[8] A Aizawa The feature quantity: an information- theoretic perspective of tfidf-like measures Proceedings of the 23rd ACM SIGIR conference on research and development in information retrieval (2000), pp. 104-111
[9] Figueira Alvaro, Oliveira Luciana.”The current state of fake news: chal- ′lenges and opportunities.” Procedia Computer Science. 121 (2017),pg 817-825.
[10] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research.
[11] Ebtihal A. Hassan, Farid Meziane. "A Survey on Automatic Fake News Identification Techniques for Online and Socially Produced Data" 2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2019

Author's Profile:

Mr.V.CHANDRASEKHAR has received him MCA degree from Sri Venkateswara University in 2001,Tirupati respectively. He is dedicated to teaching field from the last 21years. He has guided P.G students. At present he is working as Associate Professor in Audisankara College of Engineering and Technology, Gudur, Tirupati(Dt), Andhra Pradesh, India.

A.V.PAVAN KUMAR has Pursuing his MCA from Audisankara College of Engineering and Technology (AUTONOMOUS), Gudur, affiliated to JNTUA in 2022. Andhra Pradesh, India.