

Supreme court dialogue classification using machine learning models

Tomin Joseph, Vijayalakshmi Adiyillam

Department of Computer Science, Christ (Deemed to be University), Bengaluru, India

Article Info

Article history:

Received Jun 14, 2021

Revised Sep 28, 2022

Accepted Nov 3, 2022

Keywords:

Legal classification model

Logistic classifier

Machine learning

Naïve Bayes model

Natural language processing

ABSTRACT

Legal classification models help lawyers identify the relevant documents required for a study. In this study, the focus is on sentence level classification. To be more precise, the work undertaken focuses on a conversation in the supreme court between the justice and other correspondents. In the study, both the naïve Bayes classifier and logistic regression are used to classify conversations at the sentence level. The performance is measured with the help of the area under the curve score. The study found that the model that was trained on a specific case yielded better results than a model that was trained on a larger number of conversations. Case specificity is found to be more crucial in gaining better results from the classifier.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Tomin Joseph

Department of Computer Science, Christ (Deemed to be University)

Hosur road, Bengaluru-560029, Karnataka, India

Email: tomin.joseph@res.christuniversity.in

1. INTRODUCTION

Predictive coding can greatly enhance the productivity of legal study in the legal domain. The objective is to find the most relevant legal document based on a query. In the legal industry, the popular approach involves increasing the volume of training data to show efficient results [1]. Text classification is a popular study in the areas of library management, informatics, and other domains focused on natural language processing (NLP). The areas of study largely involve shorter documents in classification. However, there are studies that are focused on larger document classification techniques [2]. The approach over the years has been to reduce the massive amount of time spent finding the right document for a certain query. This is largely due to the massive cost involved in processing the data manually. Machine learning has been an acceptable approach over the years to reduce the dependence on manual approaches [3]. The reason for this continued interest in this domain is because of an exercise called “Discovery”. This process is closely related to predictive coding, but in the litigation process in courts across the United States (US), there is a need to produce relevant documents pertaining to an issue that is being addressed in court [4]. The goal was to challenge the traditional human reviewers and the algorithm has gained incredible interest from the analytics domain and was found to be effective. However, over the years there has been a considerable shift in the idea of classification. The approach has been to introduce a concept known as explainable artificial intelligence (AI). These systems can successfully locate interesting arguments within responsive documents [5]. The domain has also faced multiple difficulties including the need for annotated legal data sets which are increasingly rare because of the specificity of the language-specific domain. This can prevent researchers and practitioners from achieving the necessary efficient models. There has been significant improvement in addressing this concern by introducing automated legal data set classification [6]. Legal documents are also

known to have complex structures. This requires the introduction of novel approaches. Graph-based approaches have proven to be valuable when the document has these complex structures in place [7]. Preprocessing techniques have been deployed to extract more linguistic information from the legal text. There have been some fundamental approaches such as lemmatization and stop word techniques to improve the accuracy of classification [8]. Convolutional neural networks (CNNs) have predominantly been shown to be machine learning algorithm that is effective compared to other learning techniques such as support vector machine, logistic regression (LR) model, and random forest algorithm [9].

Classification of text documents can be done at different levels: document, paragraph, sentence, and sub-sentence level. This implies that the classification is achieved largely based on the level of the document [10]. Legal text and documents are largely considered unstructured data. The algorithms employed require these to be converted to a structured feature space for mathematical modeling which is popularly referred to as feature extraction. The term frequency-inverse document frequency (TF-IDF) denotes the importance of a word in a collection. It is popularly employed to identify the significance of a word in a document [11]. Another significant tool in feature extraction is word2vec. This tool emphasizes the need to find synonymous words which help in constructing a dictionary for the learning process [12]. An unsupervised approach which later became popular in the NLP domain was GloVe which combines the best of two model approaches [13]. Due to the increasing demand on performance, there is a need to reduce the dimensionality of the dataset further. This is popularly referred to as dimensionality reduction and is a key process in identifying significant words in a large legal document [10]. A primary classification technique that has pushed the boundaries of the classification algorithms is LR. The objective is to perform binary classification. It uses a function that takes a dependent variable as input to identify the correct class of a document [14]. The naïve Bayes (NB) classification algorithm is a successful classification algorithm based on the popular Bayes theorem where every pair of features is independent [15]. Popular techniques such as k -nearest neighbor and support vector machine are unsupervised learning techniques that have gained interest in NLP domains over the years [16]. Tree-based classifiers such as decision tree algorithms and random forest algorithms have been known to be efficient algorithms in text classification [17]. Automatic classification of text documents can also be achieved by employing graph based algorithms such as conditional random fields (CRF) which are implemented by considering the neighborhood of data items in a graph structure [18]. Lately, deep learning approaches have paved great pathways for classification algorithms. This is due to the sheer size of input data the algorithms are able to intake [19]. The final important task in text classification is to evaluate the efficiency of the model. Accuracy is a measure that has been increasingly used in text classification, however, it cannot be employed when the data set is unbalanced [20]. The F-score is a more sophisticated evaluation metric that can measure text classification with links to search engines [21]. The receiver operating characteristics (ROC) curve is a curve that is plotted by using the true positivity rate against the false positivity rate. An important piece of information that is derived from the ROC curve is the area under the curve (AUC) [22].

This study focuses on deriving the ROC curve and studying the AUC. This is done by taking two popular techniques namely the LR model and the NB model to classify legal sentences. These legal sentences are derived from dialogues in a supreme court.

2. RESEARCH METHOD

2.1. Data preparation

The data set for this work is derived from the Supreme Court Dialogs Corpus v(1.01) which was released in 2012 [23]. The corpus contains oral arguments in US Supreme Court. There are in total 50,389 conversational exchanges between 11 justices and 311 participants. The data set was prepared by manually picking the sentence from the corpus and manually annotating whether the legal sentence is from the justice or a participant. This information is already present in the data set. The justice statement is given the value 1 (i.e. justice=1) and the non-justice statement is given the value 0 (i.e. justice=0). The study is focused on creating a balanced dataset. In Figure 1, we observe that the number of justice statements is equal to the number of non-justice statements. The natural language toolkit (NLTK) is used to preprocess the existing sentences in the data set. The pre-processing step that is employed here was the removal of stop words. Words such as “the”, “a” or “an” which a search engine typically ignores were removed in this step [24].

2.2. Data exploration

The NLTK tool kit helps in the visualization of the data. The frequencies of words used by the justice and the non-justice help comprehend the difference of words used, respectively. In Figure 2, we notice that the justice statement has a large number of references to justice itself (“I”). In Figure 3, we notice a shift in the frequent words used by the non-justice including reference to the word “and”.

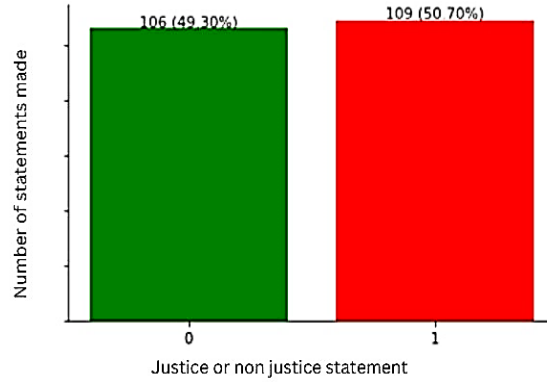


Figure 1. Data set balance

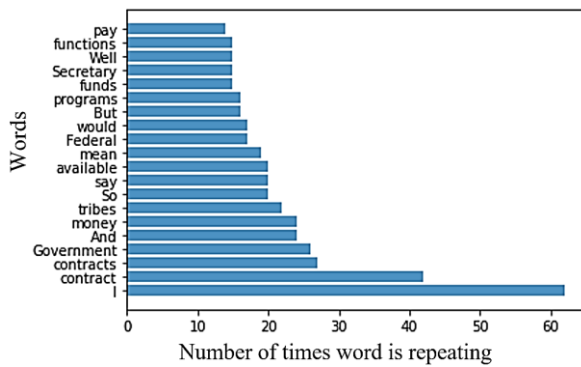


Figure 2. Frequency of words used by justice

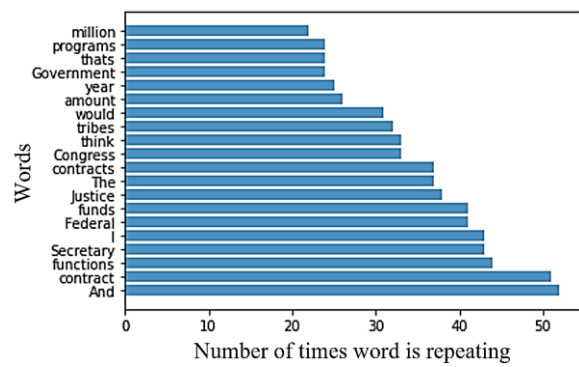


Figure 3. Frequency of words used by non-justice

2.3. Data engineering

The dimensionality reduction of the data set can be done with the help of lemmatization. Stemming or lemmatization involves combining derivationally different words into a single word [25]. This is done to increase the specificity of unique words. It will enhance the efficiency of classification when lemmatization is implemented. The feature extraction process is implemented with the help of document term frequency (DTM). In DTM, each row represents a conversation from the Supreme Court Dialog and each column represents a unique word from the conversation. This helps in analytics as the text is converted to a numeric form that can be further processed by the algorithm. Subsequently, a large number of columns would be created to accommodate lengthy documents in the process. If the word appears in the conversational log, a corresponding value of 1 is attributed to that particular word. A drawback of using DTM would be the importance given to high occurrences but less important words. This can be addressed by measuring the term frequency-inverse document frequency (TF-IDF) of a word. The TF-IDF can be measured as (1),

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right) \tag{1}$$

where $w_{x,y}$ stands for the TF-IDF weight of x , the term $tf_{x,y}$ stands for the frequency of x in y , the term df_x stands for the number of documents containing x , and N stands for the total number of documents.

2.4. Training

An LR model is built with the idea of allowing a user to make binary decisions. It is more specifically used in the case when the target variable is categorical in nature. To predict the nature of a class, we make use of threshold values. This can be decided with the help of a logistic function or Sigmoid function. LR models the probability of a certain class.

The NB classifier is based on the fundamental NB assumption that each feature makes an independent and equal contribution to the outcome. Bayes theorem finds the probability of an event occurring given that another event has already taken place. It can be mathematically represented as:

$$p(A|B) = \frac{p(B|A)P(A)}{p(B)} \tag{2}$$

$$p(y_i|x_1, x_2 \dots) = \frac{p(x_1, x_2, \dots, x_n|y_i) \cdot p(y_i)}{P(x_1, x_2, \dots, x_n)} \tag{3}$$

$P(A/B)$ gives the probability of A occurring given that B has already taken place. NB classifier calculates the probability of a class occurring given a set of feature values. Here y_i indicates the class of probability given a set of features $x_1, x_2 \dots, x_n$ occurring. NB assumes that all features are independent over each other. NB is known to work with data that has high dimensionality such as text documents. In the study, the manually annotated data set is passed to the LR model and the NB classification model. The ROC curve is drawn up. The area under the ROC curve (AUC) score is used to measure the efficiency of the model.

3. RESULTS AND DISCUSSION

There are different legal cases taken for the purpose of the study. The following sub-sections will emphasize the size of the data set for each study and will also mention the training and validation split. In this section, the result of the study is published and at the same time is given a comprehensive discussion. There are three data sets prepared for the purpose of the study: Cherokee Nation against Thompson and Thompson, Antonio Dwayne Halbert v. Michigan, and a combined data set to measure the performance of the classifiers.

3.1. Cherokee Nation against Thompson and Thompson

The data set for this particular case involves a total of 215 conversations between the justice and the corresponding lawyer representatives. The case contains a total of 106 statements made by the non-justice and 109 statements by the justice. The training data involves 193 statements and 22 statements for validation. From Figures 4 and 5, we gather the AUC scores for the data set prepared using Cherokee Nation against Thompson and Thompson. The NB classifier has an AUC score of 88.54% whereas the LR model returned an AUC score of 67.72%. Clearly, the NB classifier has outperformed LR model.

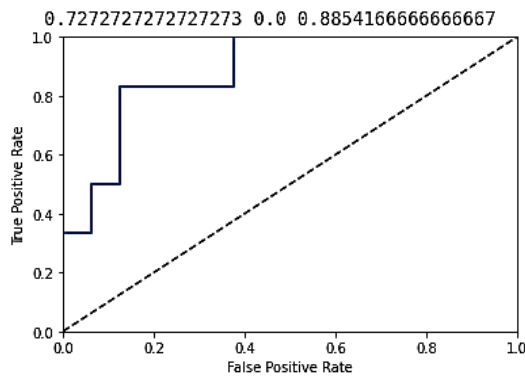


Figure 4. NB classifier using Cherokee Nation against Thompson and Thompson

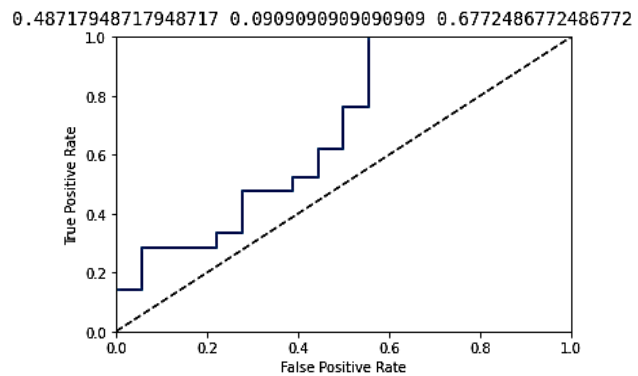


Figure 5. LR model using Cherokee Nation against Thompson and Thompson

3.2. Antonio Dwayne Halbert v. Michigan

The data set for this particular case involves a total of 174 conversations between the justice and the corresponding lawyer representatives. The case contains a total of 85 statements made by the non-justice and 89 statements by the justice. The training data involves 156 statements and 18 statements for validation. From Figures 6 and 7, we gather the AUC scores for the data set prepared using Antonio Dwayne Halbert v. Michigan. The NB classifier has an AUC score of 83.74% whereas the LR model returned an AUC score of 67.72%. Clearly, the NB classifier has outperformed LR model.

3.3. Combined cases

The data set for this particular case involves a total of 384 conversations between the justice and the corresponding lawyer representatives. The case contains a total of 191 statements made by the non-justice and 198 statements by the justice. The training data involves 350 statements and 39 statements for validation. From Figures 8 and 9, we gather the AUC scores for the data set prepared using Antonio Dwayne Halbert v.

Michigan. The NB classifier has an AUC score of 67.72% whereas the LR model returned an AUC score of 67.72%. The NB classifier performed as equally worse as the LR model.

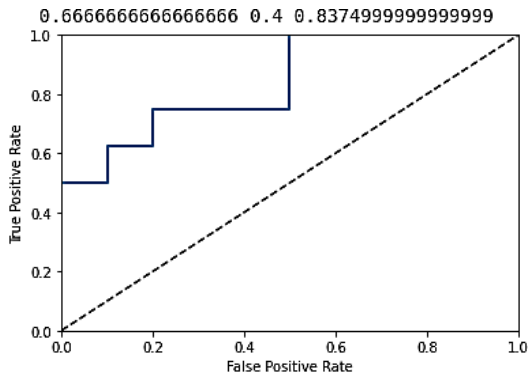


Figure 6. NB classifier using Antonio Dwayne Halbert v. Michigan

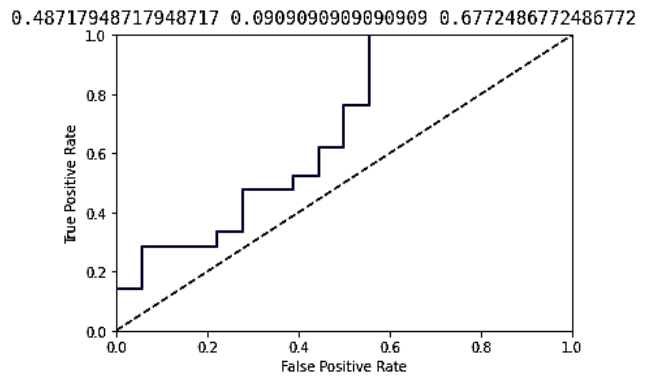


Figure 7. LR model using Antonio Dwayne Halbert v. Michigan

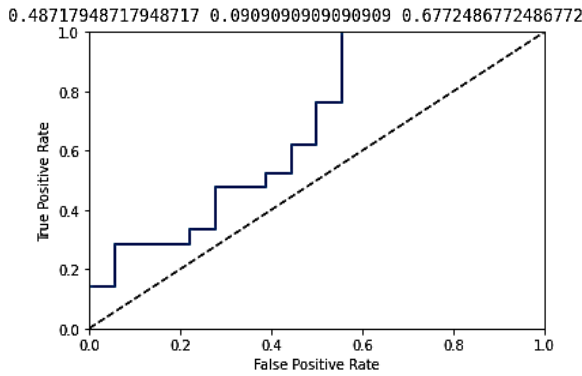


Figure 8. NB classifier using combined cases

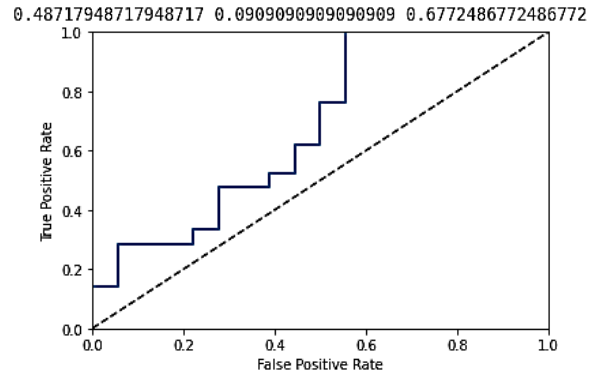


Figure 9. LR model using combined cases

4. CONCLUSION

The work carried out was to successfully classify legal conversations taking place in a supreme court using both NB classifier and LR model. The evaluation metric used was the AUC score from the ROC curve. The study conducted can successfully conclude with evidence that the NB classifier performs better than the traditional LR model. However, we notice that when cases are combined, they perform equally worse. This proves that case specificity greatly enhances the classification score. Case specificity can help classification models perform better in courts. However, a model that is trained using a large collection of conversations may not necessarily give the required performance. In the future, unsupervised modeling can be used to classify legal conversations in a supreme court.




REFERENCES

- [1] F. Wei, H. Qin, S. Ye, and H. Zhao, "Empirical study of deep learning for text classification in legal document review," in *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, pp. 3317–3320, doi: 10.1109/BigData.2018.8622157.
- [2] L. Wan, G. Papageorgiou, M. Seddon, and M. Bernardoni, "Long-length legal document classification," *arXiv: 1912.06905*, 2019.
- [3] R. Chhatwal, P. Gronvall, N. Huber-Fliflet, R. Keeling, J. Zhang, and H. Zhao, "Explainable text classification in legal document review a case study of explainable predictive coding," in *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, pp. 1905–1911, doi: 10.1109/BigData.2018.8622073.
- [4] H. L. Roitblat, A. Kershaw, and P. Oot, "Document categorization in legal electronic discovery: computer classification vs. manual review," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, pp. 70–80, Jan. 2010, doi: 10.1002/asi.21233.
- [5] C. J. Mahoney, J. Zhang, N. Huber-Fliflet, P. Gronvall, and H. Zhao, "A framework for explainable text classification in legal document review," in *2019 IEEE International Conference on Big Data (Big Data)*, Dec. 2019, pp. 1858–1867, doi: 10.1109/BigData47090.2019.9005659.




- [6] D. Song, A. Vold, K. Madan, and F. Schilder, "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training," *Information Systems*, vol. 106, May 2022, doi: 10.1016/j.is.2021.101718.
- [7] G. Li, Z. Wang, and Y. Ma, "Combining domain knowledge extraction with graph long short-term memory for learning classification of chinese legal documents," *IEEE Access*, vol. 7, pp. 139616–139627, 2019, doi: 10.1109/ACCESS.2019.2943668.
- [8] T. Gonçalves and P. Quaresma, "Is linguistic information relevant for the classification of legal texts?," 2005, doi: 10.1145/1165485.1165512.
- [9] R. Keeling *et al.*, "Empirical comparisons of CNN with other learning algorithms for text classification in legal document review," in *2019 IEEE International Conference on Big Data (Big Data)*, Dec. 2019, pp. 2038–2042, doi: 10.1109/BigData47090.2019.9006248.
- [10] Kowsari, J. Meimandi, Heidarysafa, Mendu, Barnes, and Brown, "Text classification algorithms: a survey," *Information*, vol. 10, no. 4, Apr. 2019, doi: 10.3390/info10040150.
- [11] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.
- [12] Y. Goldberg and O. Levy, "Word2Vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method," *arXiv:1402.3722*, Feb. 2014.
- [13] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
- [14] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. Wiley, 2013.
- [15] R. R. Larson, "Introduction to information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 4, pp. 852–853, 2010.
- [16] E.-H. S. Han and G. Karypis, "Centroid-based document classification: analysis and experimental results," *European conference on principles of data mining and knowledge discovery*, 2000, pp. 424–431.
- [17] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, 2001, pp. 282–289.
- [18] R. Angelova and G. Weikum, "Graph-based text classification," in *Proceedings of the 29th annual international ACM SIGIR Conference on Research and Development in Information Retrieval-SIGIR '06*, 2006, doi: 10.1145/1148170.1148254.
- [19] K. Kowsari, D. E. Brown, M. Heidarysafa, K. Jafari Meimandi, M. S. Gerber, and L. E. Barnes, "HDLTex: hierarchical deep learning for text classification," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2017, pp. 364–371, doi: 10.1109/ICMLA.2017.0-134.
- [20] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, Mar. 2005, doi: 10.1109/TKDE.2005.50.
- [21] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in *Australasian Joint Conference on Artificial Intelligence*, 2006, pp. 1015–1021.
- [22] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982, doi: 10.1148/radiology.143.1.7063747.
- [23] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg, "Echoes of power: Language effects and power differences in social interaction," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 699–708.
- [24] M. Lobur, A. Romanyuk, and M. Romanyshyn, "Using NLTK for educational and scientific purposes," in *2011 11th international conference the experience of designing and application of CAD systems in microelectronics (CADSM)*, 2011, pp. 426–428.
- [25] I. Zeroual and A. Lakhouaja, "Arabic information retrieval: stemming or lemmatization?," in *2017 Intelligent Systems and Computer Vision (ISCV)*, Apr. 2017, pp. 1–6, doi: 10.1109/ISACV.2017.8054932.

BIOGRAPHIES OF AUTHORS



Tomin Joseph    is a Ph.D. scholar from the Department of Computer Science at CHRIST (Deemed to be University) central campus. His focus area of research is currently NLP. He has a keen interest in areas related to sports analytics and teaching. He can be contacted at tomin.joseph@res.christuniversity.in.



Vijayalakshmi Adiyillam    is an assistant professor from the department of Computer Science at CHRIST (Deemed to be University) central campus. Her current focus area of research includes IoT, facial recognition systems, and NLP. Her outstanding achievement includes receiving the university-level first rank for M.Sc. Computer Science program from Mangalore University. She can be contacted at vijayalakshmi.nair@christuniversity.in.