An Empirical Comparative Study of Machine Learning Algorithms for Telugu News Classification

S.V.S Dhanush, Tsaliki Satya Ganesh Kumar,
Dharavathu Rohith, Penaka Vishnu Reddy,
K P Soman, and Sachin Kumar S*

Amrita School of Artificial Intelligence, Coimbatore,
Amrita Vishwa Vidhyapeetham, India.
s_sachinkumar@cb.amrita.edu, svsdhanush@gmail.com,
satyaganeshkumartsaliki@gmail.com , rohith.dharavathu.112@gmail.com,
vishred18@gmail.com, deanaie@amrita.edu

Abstract. Amidst escalating data growth, effective classification in diverse domains, including the news industry, is imperative. However, relying solely on human intervention for classification is unfeasible. Addressing the complexities of the Telugu language and leveraging Natural Language Processing (NLP), this study employs classification techniques. Custom Machine Learning and Deep Learning models are developed, utilizing various word embeddings, aiming to enhance accuracy and efficiency in categorizing newspaper articles. The research tackles challenges of unstructured text, attributes, NLP techniques, missing metadata, and algorithm selection. The proposed model offers both generality and efficiency, systematically classifying text documents and demonstrating significant improvements in accuracy through innovative techniques.

Keywords: Random Kitchen Sink (RKS) · Logistic regression · Multinomial Naive Bayes · Multilayer Perceptron · Natural Language Processing (NLP)· Deep Learning · Classification.

1 Introduction

As information continues to grow exponentially, the need to effectively analyze and classify large amounts of data becomes more important. Text classification is an important task in organizing and labeling documents based on their content [2]. This is achieved by building a model trained on labeled data. However, classifying documents presents many challenges, such as processing big data, that are quite difficult due to their sheer size. Additionally, high dimensionality characterized by a large number of attributes can degrade classifier performance [2]. The selection of suitable features for document representation has been found

C. Kiran Mai et al. (eds.), *Proceedings of the Fourth International Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023)*, Atlantis Highlights in Computer Sciences 18,
https://doi.org/10.2991/978-94-6463-314-6_12

to be important, and binary representations and term frequencies are common methods.

This research attempts to address these challenges by focusing on news article classification using machine learning algorithms. A comprehensive analysis and comparison of different classification algorithms is performed to determine the most appropriate approach. In addition, the effectiveness of different sets of functions is evaluated with the goal of optimizing the accuracy of news article classification.

In the field of data generation, most of the data is generated in unstructured form and includes various types such as images, web data, social media content, videos, textual information, audio data. Text classification proves to be an important application for gaining meaningful insights from this vast and heterogeneous data. Specifically, text classification is the assignment of predefined categories to a collection of text documents [11]. News headlines in particular benefit from this application as they can be automatically sorted by relevant topic to improve the efficiency, accuracy and timeliness of updates. Leverage machine learning and deep learning classifiers to minimize the need for human intervention, reduce time-consuming work, and reduce the risk of inaccurate results.

The remaining part of the paper is organized as follows. Section 2 provides the information regarding the literature review. Section 3 presents the methodology. Section 4 describes the dataset. Section 5 discusses about data preprocessing. Section 6 consists of experimental results. Finally, Section 7 offers a conclusion summarizing the key findings and contributions of the study.

2 Related Work

In a study by Veerraju et al. the classification of Telugu news headlines using a combination of Machine Learning (ML), Deep Learning (DL) and Naive Bayes (NB) classifiers[3]. They used Support Vector Machines (SVM) and Recurrent Neural Networks (RNN) models to perform classification tasks, using NB classifier they have achieved an accuracy of 90%. Sultana et al. conducted a study focused on Telugu headline classification using different ML and DL models, using Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB) algorithms to perform classification tasks. Their study achieved an accuracy rate of 79% using the SVM[13]. Kumar et al. used the DL algorithms such as LSTM, Bi-LSTM, CNN, GRU, Bi-GRU models to solve classification tasks, achieving an accuracy of 69.96% for GRU[10].

This research aims to improve the effectiveness and precision of text classification, particularly in the field of news articles, by integrating various approaches and algorithms. The proposed methods have been rigorously tested and evaluated using statistical metrics, ensuring significant improvements in the classification process.

3 Methodology

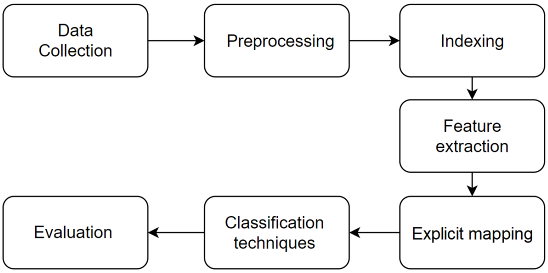

Fig. 1. Methodology

The methodology used to classify Telugu headlines involves many important steps to ensure accurate and efficient classification. The implementation process is described below.

Random Kitchen Sink (RKS): Using features taken from CountVectorizer, we improved our feature-based clustering system using the Random Kitchen Sink (RKS) method. We solved problems caused by complex hierarchical data by using RKS's ability to generate complex data representations. Improved classification accuracy and processing efficiency were achieved through projecting specialized information onto a heterogeneous subspace while preserving important data and dimensions [8]. This connection not only improves our classification process, but also transforms the dataset of Telugu news headlines for more factual information.

To define a vector transformation, the equation $\phi_\sigma(d)_i$ computes the cosine and sine of the dot product for certain vectors ω. The modified vector takes the form of a matrix after normalizing by the square root of P. This equation uses trigonometric and matrix operations to represent the vector mapping [1]

$$\phi_\sigma(d)_i = \frac{1}{\sqrt{P}} \begin{bmatrix} \cos(d_i^T)\varOmega_1 \\ \cos(d_i^T)\varOmega_2 \\ \vdots \\ \cos(d_i^T)\varOmega_P \\ \sin(d_i^T)\varOmega_1 \\ \sin(d_i^T)\varOmega_2 \\ \vdots \\ \sin(d_i^T)\varOmega_P \end{bmatrix} \tag{1}$$

Evaluation and Model Selection:

The performance of each trained model is evaluated using appropriate metrics such as precision, accuracy, recall, and F1 score. These metrics provide insight into the model's ability to accurately classify Telugu news headlines. Based on the evaluation results, the model that achieves the highest accuracy and shows good performance in classifying Telugu news headlines is considered as the most efficient model for this task. Selected models are further analyzed and validated to ensure their validity and reliability in real-life scenarios.

By incorporating feature extraction techniques using CountVectorizer, dimensionality reduction using RKS, and model training using Multinomial Naive Bayes and other ML algorithms, the proposed methodology is an accurate and efficient way to classify Telugu news headlines[12]. The purpose is to achieve classification. This research contributes to advances in natural language processing techniques in the field of text classification, with a particular focus on the Telugu domain.

3.1 Logistic Regression

Logistic regression focuses on its ”logistics” aspect, incorporating a logistic function into its classification algorithm to distinguish the category[9]. We choose polynomial logistic regression for our multilabel data. In text classification, this method identifies variables, calculates the coefficients for each input, and assigns a frame-based text type based on this data.

$$P\left(Y = \frac{1}{X}\right) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}} \tag{2}$$

P(Y = 1/X) is the probability that the outcome is 1 given the values of the independent variables (X).
β_0 is the intercept term.
β_1, β_2, ..., β_n are the coefficients of the independent variables.
X1, X2, ..., Xn are the independent variables. e is the exponential function.

3.2 Multinomial Navie Bayes

Multinomial Naive Bayes (MNB) is a popular probabilistic technique in text classification [6]. Operating on labeled datasets, MNB uses frequency estimation parameter learning, which computes matched frequencies from the data to estimate the probability of the word [10].

$$P\left(\frac{c}{X}\right) = \frac{n_c \cdot P\left(\frac{x}{c}\right)}{\sum_k n_k \cdot p(x|k)} \tag{3}$$

P(c—x) is the probability of class c given the features x.
nc is the number of samples in class c.
p(x—c) is the likelihood of the features x given class c.
n_k is the number of samples in class k.
p(x—k) is the likelihood of the features x given class k.

3.3 Support Vector Machine

The Support Vector Machine (SVM) is a supervised learning model seeking a hyperplane to separate different-class data points in a two-dimensional space [10]. It aims to maximize separation for relevant classes in an n-dimensional space, with closest points designated as Support Vectors[14]. While SVM doesn't inherently support multiclass classification, approaches like One-vs-One are used.

$$(w * x) + b = 0 \tag{4}$$

w is the weight vector perpendicular to the hyperplane.
b is the bias term.

3.4 Random Forest

The Random Forest algorithm, an ensemble learning approach, merges multiple decision trees for predictions. Each tree in the ensemble trains on a random subset of the dataset, and the final prediction results from aggregating individual tree forecasts. Thus, the Random Forest comprises these trees that jointly classify new inputs based on the provided vector [10].

3.5 Multilayer Perceptron

Input, hidden, and output layers make up the Multilayer Perceptron (MLP), a deep neural network [4]. It begins with input nodes and ends with an output layer that represents the classes ”cite” and ”inproceedings”. The output layer of our model consists of 5 nodes, each of which represents a class. Our MLP uses'relu' activation and 'adam' solver and has one hidden layer with 100 neurons. While 'adam' optimises training weight, 'Relu' incorporates non-linearity to capture complicated data patterns.

3.6 Long Short-Term Memory

LSTM (Long Short-Term Memory) is a specialized recurrent neural network architecture adept at modeling long dependencies in sequential data, suitable for time series and natural language processing [5]. We experiment with LSTM layers: one model with 1 layer (128 units) and another with 2 layers (128 units each). Inputs are reshaped into tensors. The single-layer LSTM has an output layer (5 classes) with softmax activation, compiled with sparse categorical cross-entropy and Adam optimizer. The 2-layer LSTM replicates this, adding a 128-unit layer for input transition.

3.7 Convolutional Neural Networks(1D)

We investigate several CNN layer configurations, similar to LSTM, including one model with a single CNN layer and another with two CNN layers.

Convolutional, max pooling, and dense layers with softmax activation are used for the single-layer CNN. Training is driven by sparse categorical cross-entropy loss and the Adam optimizer. Similar to this, the dual-layer CNN uses 128 and 64 filters in successive convolutional layers, followed by maximum pooling. The output is flattened, coupled to a dense layer, and the class probabilities are activated using softmax.

4 Description of the Telugu news dataset

The **Telugu Newspaper Articles Dataset** is a comprehensive collection of Telugu news articles primarily sourced from the Andhra Jyoti Newspaper website. This dataset aims to facilitate research on multi-class text classification problems, especially in the context of Telugu news articles [13].

The dataset is split into two main files. A training set with 17,306 unique entries and a test set with 4,329 unique entries [13]. The dataset has five attributes namely Si.no which acts as the unique identifier for each news article and the published date (Date), heading of the article (Heading), the content of the news article (Body) and the target label Category represents to which category the news article belongs to. The categories in the dataset includes business, editorial, entertainment, national, and sports.

These categories provide an overview of the types of news covered by the dataset, from business news to sports news. These different categories allows us to perform different multiclass classification tasks on this data set.

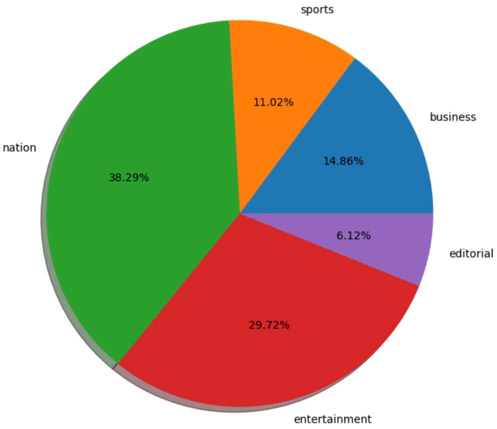

Fig. 2. Percentage Distribution of Classes

5 Data Preprocessing

Data preprocessing was performed on the dataset to handle null values and eliminate special characters. The preprocessing step focused on selecting two specific columns from the dataset, namely "heading" and "topic." Additionally, the labels were converted into categorical codes as part of the preprocessing procedure [11].

Random Kitchen Sink: Random Kitchen Sink (RKS) is a powerful technique employed in NLP for explicit feature mapping and dimensionality expansion. It leverages the generation of random features to approximate the kernel trick, offering computational efficiency advantages over kernel methods like SVM. To maximize the potential of RKS, it is advisable to incorporate techniques such as feature selection, dimensionality reduction (e.g., PCA), and hyperparameter tuning[7]. It is crucial to thoroughly evaluate RKS's performance in comparison to other techniques such as linear models or deep learning, and also to explore alternative methods like word embeddings. The effectiveness of RKS heavily relies on the specific NLP task and dataset at hand, necessitating meticulous experimentation and evaluation for optimal results.

5.1 Classification Techniques

After training a custom word embedding model specialized for Telugu texts (excluding GloVe), the resulting word embeddings can be used as input for classification algorithms. The classification algorithms used in this study are Support Vector Machine with Linear Kernel, Support Vector Machine with Polynomial Kernel, Logistic Regression, Multinomial Naive Bayes, Random Forest, K Nearest Neighbors (KNN), and Multilayer Perceptron (MLP) and long-short-term memory (LSTM) and also Convolutional Neural Network (CNN).

Employing these diverse classification algorithms, both traditional machine learning and deep learning, we are able to perform comprehensive analysis of Telugu text data, allowing us to more thoroughly understand the underlying patterns and structures.

6 Experimental Results and Discussions

Data acquisition and preprocessing: A large collection of Telugu news headlines collected in preliminary stage. To improve their quality and usefulness for classification, these data can be carefully prepared. There are different stages in this procedure. First, the text is converted to Word format and then segmentation is performed, which separates the data records into individual words. Stop word removal is used to remove semantically frequently meaningless terms such as articles, prepositions, and pronouns to minimize noise and improve relevancy. In addition, word root forms were identified using radical methods, which encouraged generalization and minimized diversity in the data set.

Feature extraction with CountVectorizer: The dataset is feature extracted using CountVectorizer technique after data preparation. Treating each text document as an unordered set of words, it converts the text into a matrix. The frequency of occurrence of words in each document is calculated, creating a numerical representation of the data set. By accurately capturing the meaning of words, this technique improves future classification.

Result analysis is used to evaluate the results of the classifier as given in Table 1. Accuracy signifies the rate at which Telugu News Headlines are correctly classified. Precision indicates the trustworthiness of our system's classification predictions. Recall represents the share of relevant articles accurately predicted and categorized by our model. F1 score states that harmonic mean of the precision and recall. Confusion matrix measures classification.

Table 1. Performance of different models.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	92.3%	91%	88%	89%
Multinomial NB	94%	89%	81%	84%
SVM	88.25%	89%	81%	84%
Random Forest	65%	66%	59%	58%
KNN	52%	65%	54%	53%
Multi-Layer Perceptron	92%	94%	88%	90%
LSTM 1-Layer	84%	83%	83%	80%
LSTM 2-Layer	94%	94%	91%	92%
CNN 1-Layer	92%	92%	89%	90%
CNN 2-Layer	91%	89%	90%	89%

According to our investigation of multiple machine learning models for categorising news headlines, Multinomial Naive Bayes and Logistic Regression had an accuracy rates of 94% and 92.3% respectively, and were thus trustworthy options while we need to choose a ML algorithm. With accuracy rates of 88.25% and 84% , respectively, the Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) with 1 layer models also performed well but LSTM with 2 layers has given us more promising results of 94% when compared to other neural networks such as CNN with 1 layer gave us 92% where as CNN with 2 layers gave us 91% accuracy. The Multinomial Naive Bayes and LSTM with 2 layers demonstrated accuracy of 94%, displaying their applicability to classification jobs for news headlines. Overall, the best models can be used to build reliable categorization systems while other models can still be improved upon.

While Multi-Layer Perceptron and LSTM 2-Layer gave best precision result of 94%. While LSTM 2-Layer and CNN 2-Layer gave best recall results of 91%

and 90% respectively. LSTM 2-Layer gave best F1 score of 92%. We used a learning rate of 5 and a batch size of 433 for the DL models we implemented.

The following are the confusion matrices for few models that we classified. Figure 3 represents the confusion matrix for Multinomial Naive Bayes. Similarly figures 4 and 5 represents confusion matrix for KNN and Random Forest and respectively.

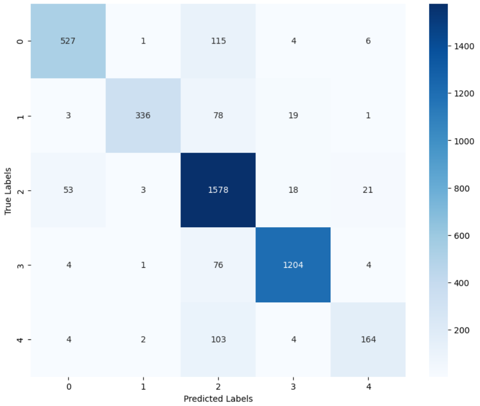

Fig. 3. Multinomial Naive Bayes Confusion matrix

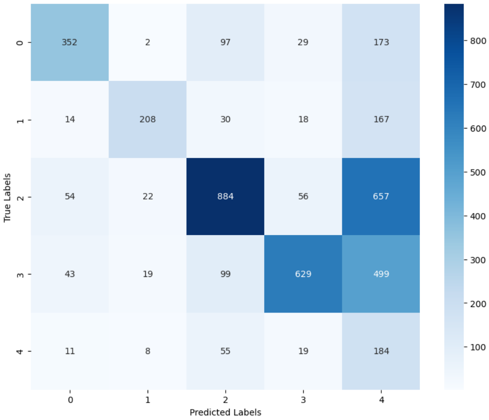

Fig. 4. KNN Confusion matrix

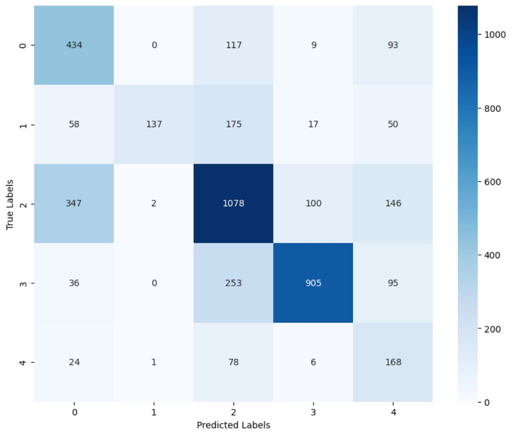

Fig. 5. Random Forest Confusion matrix

7 Conclusion

This research study focused on using multiple machine learning and deep learning models to classify new articles, and extensive evaluations were conducted to assess message performance. The multinomial naive Bayes and LSTM with 2 layers model has an accuracy rate of 94%, making it a reliable option for telugu news article classification. Similarly, the logistic regression model showed a laudable accuracy of 92.3%.

Additionally, the support vector machine (SVM) model achieved an accuracy of 88.25%, demonstrating its effectiveness in accurately classifying news articles. A long short-term memory (LSTM) model also provided satisfactory results with an accuracy of 84%. Moreover, the multi-layer perceptron model showed a respectable accuracy of 92%, further supporting its suitability for the new article classification task. These results highlight the applicability and feasibility of these models for building reliable text classification systems.

In summary, this study provides valuable insight into the effectiveness of different machine learning and deep learning models in telugu new article classification, with Multinomial Naive Bayes and LSTM models standing out as particularly reliable options. This result highlights the potential for these models to be effectively used for message classification tasks in real-world applications.

References

1. Athira, S., Harikumar, K., Sowmya, V., Soman, K.: Parameter analysis of random kitchen sink algorithm. International Journal of Applied Engineering Research **10**(20), 19351–19355 (2015)
2. Dhar, A., Mukherjee, H., Dash, N.S., Roy, K.: Text categorization: past and present. Artificial Intelligence Review **54**, 3007–3054 (2021)

3. Gampala, V., Vallapuneni, J., Kumar, A., Indurthi, R., Nichenametla, R.: Comparative study on telugu text classification using machine learning and deep learning models. pp. 1393–1398 (06 2021). https://doi.org/10.1109/ICOEI51242.2021.9453040
4. Jehad, R., Yousif, S.A.: Classification of fake news using multi-layer perceptron. In: AIP Conference Proceedings. vol. 2334, p. 070004. AIP Publishing LLC (2021)
5. Khuntia, M., Gupta, D.: Indian news headlines classification using word embedding techniques and lstm model. Procedia Computer Science **218**, 899–907 (2023)
6. Kibriya, A.M., Frank, E., Pfahringer, B., Holmes, G.: Multinomial naive bayes for text categorization revisited. In: AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17. pp. 488–499. Springer (2005)
7. Prasanth, S., Raj, R.A., Adhithan, P., Premjith, B., Kp, S.: Cen-tamil@ dravidianlangtech-acl2022: Abusive comment detection in tamil using tf-idf and random kitchen sink algorithm. In: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages. pp. 70–74 (2022)
8. Sathyan, D., Anand, K.B., Prakash, A.J., Premjith, B.: Modeling the fresh and hardened stage properties of self-compacting concrete using random kitchen sink algorithm. International journal of concrete structures and materials **12**, 1–10 (2018)
9. Shah, K., Patel, H., Sanghvi, D., Shah, M.: A comparative analysis of logistic regression, random forest and knn models for the text classification. Augmented Human Research **5**, 1–16 (2020)
10. Sri Sravya, V., Kumar S, S., Soman, K.P.: Text categorization of telugu news headlines. In: 2022 2nd International Conference on Intelligent Technologies (CONIT). pp. 1–6 (2022). https://doi.org/10.1109/CONIT55038.2022.9847875
11. Sudha, D.N., et al.: Semi supervised multi text classifications for telugu documents. Turkish Journal of Computer and Mathematics Education (TURCOMAT) **12**(12), 644–648 (2021)
12. Sultana, J., Macigi, U.R., Priya, G.: Telugu News Data Classification using Machine Learning Approach, pp. 181–194 (09 2021). https://doi.org/10.4018/978-1-7998-7685-4.ch014
13. Sultana, J., et al.: Telugu news data classification using machine learning approach. In: Handbook of Research on Advances in Data Analytics and Complex Communication Networks, pp. 181–194. IGI Global (2022)
14. Ukey, K., Alvi, A.: Text classification using support vector machine. International Journal of Engineering and Technology (IJERT) (2012)