# Multilingual Sentiment Analysis using Supervised Machine Learning Algorithms

**Shubham Ojha**
Student / VIT Chennai
shubhamojha2109@gmail.com

**Rajalakshmi R**
Prof / SCOPE, VIT Chennai
rajalakshmi.r@vit.ac.in

## Abstract

This paper utilizes Supervised Machine Learning algorithms to categorize the Malayalam language into two tasks. Task 1, involves the classification of social media texts from platforms such as Twitter, Facebook, and YouTube as either original or fake news/comments. Task 2 is dedicated to the detection and categorization of fake news in Malayalam-language articles into five categories: False, Half True, Mostly False, Partly False, and Mostly True. The shared objective is to address the challenge of accurate misinformation detection in the era of information overload, fostering trustworthy communication.(2)

## 1 Introduction

The ability to distinguish between accurate and false information has become increasingly important in a time when information is abundant and content is widely shared on digital platforms. The proliferation of false information on the internet poses a serious threat to the credibility of material and can lead to the spread of false information that can be harmful to both individuals and communities. This essay tackles the crucial task of differentiating between fake and real content, clarifying the complex problems related to the spread of false information.

The insights obtained from a large dataset obtained as part of the CodaLab Fake News Detection in Dravidian Languages competition (DravidianLangTech@EACL 2024)(1) are applied to our inquiry. This research focuses on two types of data: first, it investigates the realm of YouTube comments, using them as a unique means of assessing the veracity of content; second, it explores the complex field of fake news detection, which is divided into five categories: False, Half True, Mostly False, Partly False, and Mostly True.

We use a variety of supervised machine learning methods, such as Random Forest, SVM, Logistic Regression, and Naive Bayes, to traverse this challenging terrain. The combination of the linguistic diversity among the Dravidian languages and the complexities of content deception offers a demanding yet intriguing context for our study. This research intends to provide significant insights into the ongoing discussion on the integrity of digital information by analyzing the subtleties of fake news detection, opening the door for more reliable and efficient content authentication systems.

## 2 Dataset Discription

**Malayalam Comments**

Text: Text in Malayalam language.

label: Info about fake or original.

**Malayalam News**

News: News in the Malayalam language.

Label: This column categorizes news in 5 categories i.e. False, Half True, Mostly False, Partly False, and Mostly True.

## 3 Proposed Work

**Malayalam Comment Classification**

**Data Preparation:** Import the necessary libraries, including scikit-learn and pandas. Load the training dataset for model training.

**Vectorization of Text:** To convert the textual data into numerical vectors, use the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer.

**The Logistic Regression Model:** To find the ideal regularisation parameter (C), use cross-validation and Logistic Regression using a hyperparameter grid search. Utilize the determined ideal hyperparameters to train the classifier.

**The Random Forest Model:** To determine the optimal set of hyperparameters, such as the number of estimators, maximum depth, minimum samples split, and minimum samples leaf, use Random Forest in conjunction with a grid search. Utilize the

optimized hyperparameters to train the Random Forest model.

**The Support Vector Model(SVM):** Use the Support Vector Machine (SVM) technique for categorization. Utilizing the TF-IDF vectorized training data, train the SVM model.

**The Naive Bayes Model:** Multinomial Naive Bayes should be used for categorization. Utilizing the TF-IDF vectorized training data, train the Naive Bayes model.

**Verification Assessment of the dataset:** To evaluate the model, load the validation dataset. Take care of missing values and vectorize the text column. Evaluate each classifier's accuracy score on the validation set.

**Ideal Model Choice:** Determine which classifier on the validation set has the best accuracy. Declare this classifier to be the best model to test further. Evaluation of Testing Dataset.

**Testing:** Load the testing dataset. Utilizing the TF-IDF vectorizer, vectorize the text column of testing data. To predict labels for the test dataset, use the best classifier available.

**Conclusion:** Conclude the study by summarizing the chosen classifier's performance on the testing dataset.

**Malayalam News Classification**

**Data Preparation:** Import the necessary libraries, including scikit-learn and pandas. Load the training dataset for model training.

**Vectorization of Text:** To convert the textual data into numerical vectors, use the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer.

**Applied ML algorithms:** Similar to section 3.1, all classifiers are trained and the best model is found out for News Classfication.

**Testing:** Load the testing dataset. Utilizing the TF-IDF vectorizer, vectorize the text column of testing data. To predict labels for the test dataset, use the best classifier available.

**Conclusion:** Conclude the study by summarizing the chosen classifier's performance on the testing dataset.

The combined flowchart for both tasks is shown in Fig.1.

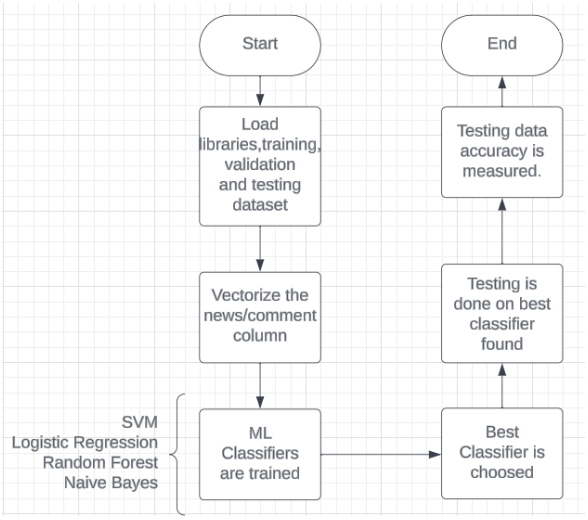


Figure 1: Combined Flowchart for both the tasks.

## 4 Models

**TF-IDF**

In natural language processing, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer is a numerical representation method.(7) It gauges a word's significance in a document in relation to a corpus. TF-IDF is a good fit for sentiment analysis in Malayalam and other multilingual contexts because it captures the distinctiveness of phrases, which aids machine learning algorithms in identifying sentiment-related features in a variety of languages.

**Support Vector Machine**

For categorization problems, a machine learning model called a Support Vector Machine (SVM) is employed.(3) It operates by locating the hyperplane in the feature space that best divides various classes. By mapping data points to a high-dimensional space and identifying the ideal hyperplane, SVM carries out categorization. Since the Radial Basis Function (RBF) kernel handles non-linear correlations in the data efficiently, it is appropriate for sentiment analysis in Malayalam and other multilingual languages. SVM is useful for sentiment analysis in a variety of languages, including Malayalam, where sentiment expressions may have non-linear structures. This is because the RBF kernel enables SVM to capture complicated patterns.

**Logistic Regression**

A popular statistical model for applications involving binary classification is logistic regression.(5) During the training phase, the choice of solvers—"liblinear," "lbfgs," "newton-cg," and "sag"—plays a critical role in optimization. While 'sag' is especially effective for large datasets, 'lbfgs' and 'newton-cg' show effectiveness in multiclass settings, and 'liblinear' is well-suited for small to

medium datasets.

'Liblinear' is frequently preferred in the context of two-category comment categorization since it is compatible with binary outcomes and computationally efficient on smaller datasets. 'lbfgs' or 'newton-cg', however, would be better appropriate for news classification tasks with five categories. These solvers are very good at multiclass scenarios; they provide better model accuracy when a larger variety of categories are present in the classification problem.

**Random Forest**

One popular ensemble learning technique for tasks involving regression and classification is called Random Forest.(4) To increase accuracy and reduce overfitting, it constructs several decision trees during training and combines their predictions. Among the crucial variables are:

n_estimators: The number of trees in the forest. max_depth: Every tree's maximum depth. min_samples_split: The bare minimum of samples needed to separate an internal node. min_samples_leaf: The bare minimum of samples necessary for a leaf node to exist. Because Random Forest can manage non-linear relationships and capture complicated patterns, it is a good choice for multilingual sentiment analysis. Because of its ensemble structure and strong parameter adjustment, it can adjust well to a wide range of linguistic traits, making it a flexible and useful method for sentiment analysis in many languages.

**Naive Bayes**

Assuming feature independence, the Naive Bayes model is a probabilistic classification algorithm built on the foundation of Bayes' theorem.(6) It is frequently employed in text classification. There are other varieties, such as Multinomial NB for discrete data like word counts and Gaussian NB for continuous data. Due to its ability to handle discrete characteristics such as word frequencies in text well and its multilingual adaptability, Multinomial NB is a good choice for multilingual sentiment analysis. Its adaptability for sentiment analysis in a variety of linguistic situations stems from its ease of use and efficiency in handling sparse and high-dimensional data.

Fig 2 Shows the different parameters used for classifiers.


Figure 2: Parameter used in grid search and Models

## 5 Result

### Malayalam Comment Classification

The study's findings provide interesting new information on how different machine learning classifiers perform when used to identify fake Malayalam comments. During training, a Naive Bayes classifier performed exceptionally well, obtaining an impressive accuracy of 78.89% on the validation dataset. On the other hand, the SVM Classifier performed somewhat better, with an accuracy of 50.01% on the validation dataset. With grid Search CV optimization, the Random Forest Classifier demonstrated a competitive accuracy of 75.5% on the validation set. Concurrently, the Logistic Regression Classifier—which was set up using grid Search CV's ideal settings and liblinear as the solver—achieved an impressive accuracy of 78.56%.

The testing dataset was then subjected to the Naive Bayes Classifier, which produced a remarkable accuracy of 78.5%. This classifier had the greatest accuracy on the validation set. These results highlight Naive Bayes's effectiveness in identifying fake Malayalam comments and highlight its potential for practical use.

| Models | Validation Data Accuracy |
|---|---|
| SVM | 50.18% |
| Random Forest | 75.58% |
| Logistic Regression | 78.65% |
| Naive Bayes | 78.90% |

Table 1: Malayalam Comment Classification Validation Data Accuracy

### Malayalam News Classification

| Models | Validation Data Accuracy |
|---|---|
| SVM | 67.55% |
| Random Forest | 63.46% |
| Logistic Regression | 65.68% |
| Naive Bayes | 61.13% |

Table 2: Malayalam Fake News Classification Validation Data Accuracy

| Classification Task | Best Model |
|---|---|
| Task 1 | Naive Bayes (78.5%) |
| Task 2 | SVM (64.0%) |

Table 3: Summary of Testing Data Accuracy

A training dataset that was divided into a 7:3 ratio was used in the study for both training and validation. For analysis, a number of classifiers were developed. With a validation accuracy of 61.12%, the Naive Bayes classifier performed worse than the SVM classifier, which performed better with an accuracy of 67.54%. A 63.45% accuracy rate was achieved by using a Random Forest Classifier with optimized parameters found by grid search CV. Similar to this, the Logistic Regression Classifier obtained an accuracy of 65.67% on the validation set by using the optimal grid search CV settings with the solver configuration set to liblinear.

The SVM Classifier was used for the testing dataset due to its better performance on the validation set, and it produced a 64% accuracy rate. These results highlight the SVM Classifier's efficacy in the context of identifying false news, offering insightful information for further study and useful applications in content authentication.

## 6 Conclusion

To sum up, this study effectively tackled two important issues related to Malayalam language processing. First, the study employed the Naive Bayes classifier to classify YouTube comments as authentic or fraudulent, and it did so with an amazing 78.5% accuracy rate. Second, the study examined how to categorize Malayalam fake news into five different groups: False, Half True, Mostly False, Partly False, and Mostly True. Using the Support Vector Machine (SVM) classifier, this classification was accomplished with a 64% accuracy rate.

Naive Bayes is preferred because of its ease of use, efficacy, and efficiency when processing linguistic data, which makes it a good fit for the task of classifying comments on YouTube. However, SVM's popularity is said to have stemmed from its capacity to manage high-dimensional data and non-linear correlations, which helped with the challenging task of categorizing bogus news into several groups. Together, the subtle advantages of SVM and Naive Bayes add to the overall success of this study project.

## References


[1] Subramanian, Malliga and Chakravarthi, Bharathi Raja and Shanmugavadivel, Kogilavani and Pandiyan, Santhiya and Kumaresan, Prasanna Kumar and Palani, Balasubramanian and Singh, Muskaan and Raja, Sandhiya and Vanaja and S, Mithunajha *Overview of the Shared Task on Fake News Detection from Social Media Text*. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, September 2023, Varna, Bulgaria. Published by Recent Advances in Natural Language Processing.

[2] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2), 1-135.

[3] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

[4] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

[5] Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley Sons.

[6] Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*.

[7] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.