Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Identification of language from multi-lingual dataset using classification algorithms

To cite this article: N Abinaya *et al* 2023 *J. Phys.: Conf. Ser.* **2664** 012009

Identification of language from multi-lingual dataset using classification algorithms

N Abinaya[1,2], P Jayadharshini[1], S Priyanka[1], S Keerthika[1] and S Santhiya[1]

[1]Assistant Professor, Department of Artificial Intelligence, Kongu Engineering College, Perundurai, India.

[2]abi9106@gmail.com

Abstract. The process of automatically identifying the language used in a text or document is known as language identification. Language identification might represent a critical step in Natural Language Processing (NLP). It entails making an effort to foresee a text's natural language. Before any actions can be conducted, it is crucial to understand the language of the text. We must build a model that can anticipate the given language using the text as a guide. This provides an answer for many computational linguists and Artificial Intelligence (AI) applications. In this study, the language in the provided text was identified using machine learning algorithms and vectorization techniques. The performance of different classification algorithms like Naïve bayes, Logistic Regression, Decision Tree and Random Forest have been compared and analyzed. Vectorization technique has been done to convert the text into matrix. This paper presents the comparison of all the above algorithms performed through various measures.

Keywords: language identification, decision tree, random forest, logistic regression, and naive bayes.

1. Introduction

There are many instances and situations where it is unknown what language a certain piece of text was originally written in. The text in real-time data does not belong to a single language. People used to combine two different languages when giving comments or remarks. Code-Mixed Data is the name given to these data. Language identification in this type of data is significant. Prior to doing other information retrieval tasks, such as machine translation, determining the language becomes essential. In these situations, automatically identifying the languages through computational approaches might be quite helpful [1]. Language identification refers to this automatic method of identifying the language used in the text. The Natural Language Processing industry has been impacted by the constant growth of data and it is a domain to become widely desired. This work aims to compare and contrast various classification models, including Naive Bayes, Logistic Regression, Decision Tree and Random Forest.

2. Related work

Anushri et al. combined six models in their work in 2022 and discovered that logistic regression and Term Frequency - Inverse Document Frequency (TF-IDF) had the best accuracy (98.45%), while

Naive Bayes model and TF-IDF had the lowest accuracy (81.33%) [1]. Language mixing is a significant problem for the Dravidian languages dataset. A research initiative on multilingual text classification was put forth in 2021. They carried out a comparison of a model with different weight values. The model performs to be good when the weight value is 1.5 [2]. Most social media users attempt to format their comments in Code-Mixed format. In these circumstances, code-mixed text plays a key part in language identification. They carried out a word-level linguistic identification [3]. Deep learning models were proposed by Jerin Mahibha et al. to be used to identify offensive language. With regard to distinguishing between offensive and non-offensive data, this study has an F1 Score of 0.865 [4]. In 2020, the identification of spoken languages on the Indic TTS dataset, developed by IIT-Madras with an accuracy of 92.35 percent, and the Indic Speech database, developed by IIIT-Hyderabad, with 100% accuracy was received [5]. An ensemble model has been developed with Naive Bayes and Support Vector Machine (SVM) on the social media text to detect offensive languages and received accuracy of 98% [6]. Automatic language and speech identification model was implemented using pre-trained models and it showed better performance than non-pretrained models [7]. Language Identification at word level is essential for code-mixed data. Different classifiers were used for identifying usage of multiple languages in social media text at word level. Among those classifiers SVM produced highest accuracy [8]. Long Short-Term Memory (LSTM) and word vectors representations are used for identifying the similar languages [9].

3. Data description

There were 17 different language datasets overall, of which 4 are Indian languages. The dataset contains 10337 instances and 2 attributes [10]. Figure 1 gives the percentage of data available for training in each language.

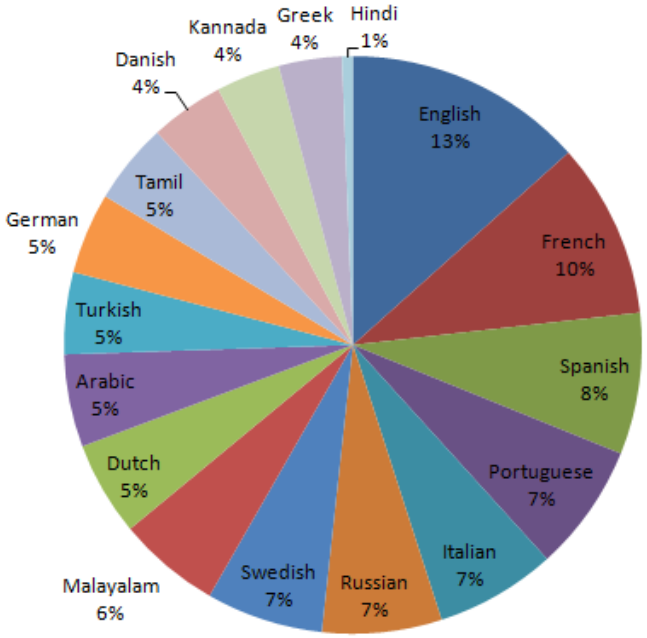

Figure 1. Training data distribution.

4. Proposed model

The suggested model for this study is depicted in figure 2. As the initial step in converting the text into numerical values for processing, the dataset of both training and test will be label encoded. The label encoder will convert the given text into numerical representation for each instance. This data is then given to the model, which is trained with various classifiers for comparison. The test data will also, after label encoding, be given to model for the prediction.

5. Performance evaluation

5.1. Naive bayes algorithm

One of the most efficient basic supervised and probabilistic machine learning techniques is the Naive Bayes algorithm. As a result, it is a probabilistic classifier which makes predictions on the likelihood of objects. It is assumed that each occurrence of a feature occurs independently of all other occurrences of features. The Bayes theorem, that states the following equations, sums everything up.

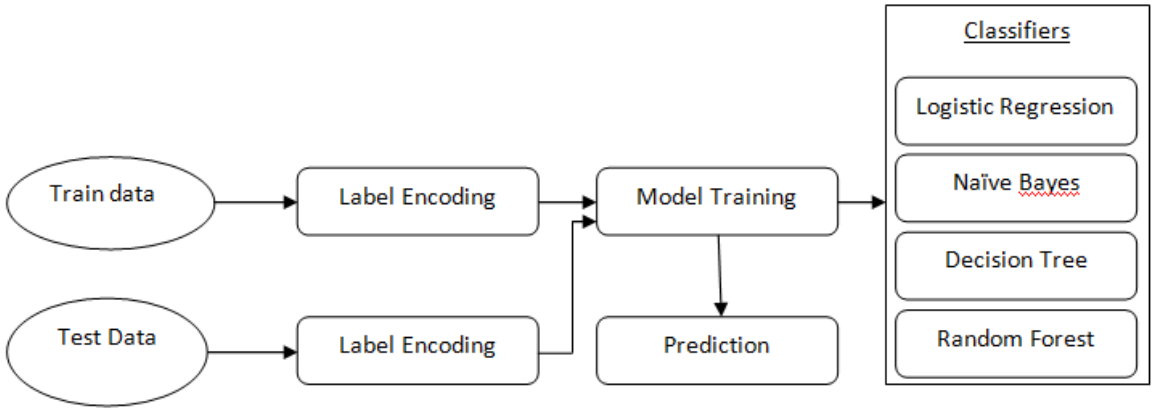

Figure 2. Proposed model.

Table 1 lists the various metrics that include Precision, Recall and F1 Score, for each of the 17 test classes that the Naive Bayes classifier identified. On every class, this model proved well.

Table 1. Performance measure of Naïve Bayes algorithm.

Class	Precision	Recall	F1-Score
Arabic	1.00	0.98	0.99
Danish	0.94	0.96	0.95
Dutch	0.96	0.96	0.96
English	0.99	0.99	0.99
French	0.99	0.98	0.99
German	0.93	1.00	0.96
Greek	1.00	0.97	0.98
Hindi	0.94	0.94	0.94
Italian	0.96	0.96	0.96
Kannada	0.87	0.99	0.93
Malayalam	0.98	0.99	0.98
Portuguese	0.98	0.98	0.98
Russian	1.00	0.99	1.00
Spanish	0.99	0.94	0.97
Swedish	0.97	0.97	0.97
Tamil	1.00	0.98	0.99
Turkish	1.00	0.97	0.99

5.2. Decision trees

A decision tree is a structure in which every single leaf node represents a class label, each internal node acts as a "test" on an feature, and every leaf displays the act outcome. Classification rules are represented by the routes from root to leaf. Expected values of competing alternatives are calculated using as a visual and analytical decision-assistance tool in decision analysis, a decision tree and the

strongly linked influence diagram are used. The various metrics assessed for the decision tree model are listed in Table 2. Compared to other models, this one performs quite poorly.

Table 2. Performance measure of decision tree.

Class	Precision	Recall	F1-Score
Arabic	1.00	0.88	0.93
Danish	0.76	0.78	0.77
Dutch	0.84	0.86	0.85
English	0.94	0.93	0.93
French	0.85	0.85	0.85
German	0.98	0.77	0.86
Greek	1.00	0.94	0.97
Hindi	1.00	0.75	0.86
Italian	0.83	0.77	0.80
Kannada	0.45	0.99	0.62
Malayalam	1.00	0.96	0.98
Portuguese	0.90	0.83	0.87
Russian	1.00	0.85	0.92
Spanish	0.75	0.75	0.75
Swedish	0.93	0.83	0.88
Tamil	1.00	0.96	0.98
Turkish	0.74	0.76	0.75

Table 3. Performance measure of random forest.

Class	Precision	Recall	F1-Score
Arabic	1.00	0.88	0.93
Danish	0.87	0.84	0.86
Dutch	0.97	0.94	0.95
English	0.97	0.97	0.97
French	0.97	0.95	0.96
German	0.99	0.91	0.95
Greek	1.00	0.95	0.98
Hindi	1.00	0.69	0.81
Italian	0.91	0.88	0.89
Kannada	0.51	0.99	0.67
Malayalam	1.00	0.96	0.98
Portuguese	0.95	0.94	0.95
Russian	1.00	0.93	0.96
Spanish	0.89	0.90	0.90
Swedish	0.93	0.91	0.92
Tamil	1.00	0.98	0.99
Turkish	0.97	0.85	0.90

5.3. Random forest

In the training stage of the random forests or random choice forests ensemble learning approach, which is used for regression, classification and other tasks, several decision trees are constructed. Random forests counteract the propensity of decision trees to overfit their training set. Although they typically outperform decision trees, gradient enhanced trees are more precise than random forests. Table 3 performs more effectively than the decision tree model and has a 92% accuracy rate.

5.4. Logistic regression

One of the most often used Machine Learning techniques is logistic regression, which is categorised under supervised learning approaches. It is a method to determine a categorical dependent variable from a set of independent factors. Regression challenges are solved using linear regression, and classification problems are solved using logistic regression. In logistic regression, a logistic function with a "S" shape is produced instead of a regression line, and the maximum values are 0 and 1. The curve of the logistic function represents the probability of anything occurring. It may produce probabilities and categorize new data using continuous and discrete data. The best categorization scheme can be simply determined using logistic regression. Table 4 displays the performance of distinct classes. It is a better model used to categorize the different classes, scoring around 95% accuracy.

Table 4. Performance measure of Random Forest.

Class	Precision	Recall	F1-Score
Arabic	1.00	0.93	0.96
Danish	0.93	0.88	0.91
Dutch	0.98	0.94	0.96
English	0.98	0.99	0.99
French	0.98	0.97	0.97
German	1.00	0.94	0.97
Greek	1.00	0.97	0.98
Hindi	1.00	0.62	0.77
Italian	0.95	0.94	0.95
Kannada	1.00	0.95	0.98
Malayalam	1.00	0.99	1.00
Portuguese	0.97	0.97	0.97
Russian	0.75	1.00	0.86
Spanish	0.93	0.94	0.93
Swedish	0.95	0.93	0.94
Tamil	1.00	0.96	0.98
Turkish	0.94	0.91	0.93

Figure 3 gives the overall accuracy of all four models. In this text classification, Naive Bayes consistently proves to be a stronger model, leading to more accurate results. Among the four classification models, the model was developed utilizing Naïve Bayes shown to produce the best accuracy of 98%, logistic regression, which gave 95% of accuracy and Random forest with 92%. The least accuracy was 85%, given by Decision tree.

6. Conclusion

Naïve Bayes classifier gives better accuracy among other models with 98%. This model can be further improved by training with more instances. The dataset used for the present study has various numbers of instances for each class. Few classes like Hindi possess only 1% of training data. So, the number of instances can be improved and made equal for all the class which may again improve the performance of the model.

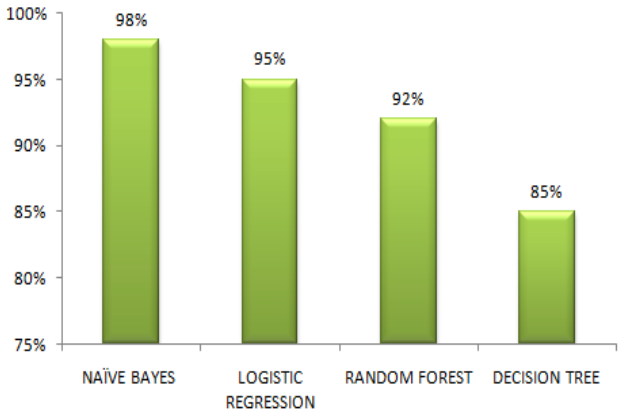

Figure 3. Overall accuracy for each model.

Acknowledgments
We would like to express our gratitude to Pavithra E, Archanaa S, Dharunraja Sr, Vignesh R, Department of Artificial Intelligence and Machine Learning, Kongu Engineering College for their role in developing the proposed model.

References
[1] Anushri Bhansali, Amit Chandravadiya, Brijeshkumar Y. Panchal, Mohammed Husain Bohara and Amit Ganatra 2022 Language Identification Using Combination of Machine Learning Algorithms and Vectorization Techniques International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) pp 1329-1334
[2] Xiaotian Lin, Nankai Lin, Kanoksak Wattanachote, Shengyi Jiang and ianxi Wang 2021 Multilingual Text Classification for Dravidian Languages
[3] S. Thara and Prabaharan Poornachandran 2021 Transformer Based Language Identification for Malayalam-English Code-Mixed Text IEEE Access vol 9 pp 118837-18850
[4] C Jerin Mahibha, Sampath Kayalvizhi, Durairaj Thenmozhi and S Arunima 2021 Offensive Language Identification using Machine Learning and Deep Learning Techniques FIRE 2021 : Forum for Information Retrieval Evaluation Working Notes, CEUR Workshop Proceedings pp 705-713
[5] Aankit Das, Samarpan Guha, Pawan Kumar Singh, Ali Ahmadian and Norazak Senu 2020 A Hybrid Meta-Heuristic Feature Selection Method for Identification of Indian Spoken Languages From Audio Signals IEEE Access vol 8 pp 181432-181449
[6] M Anand, Kishan Bhushan Sahay, Mohammed Altaf Ahmed, Daniyar Sultan, Radha Raman Chandan and Bharat Singh 2023 Deep Learning And Natural Language Processing In Computation For Offensive Language Detection in Online Social Networks by Feature Selection and Ensemble Classification Techniques Theoretical Computer Science vol 943 pp 203-218
[7] Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi and Takahiro Shinozaki 2020 Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning INTERSPEECH 2020 pp 1037 – 1041
[8] Kasthuri Shanmugalingam, Sagara Sumathipala and Chinthaka Premachandra 2018 Word Level Language Identification of Code Mixing Text in Social Media using NLP 3rd International Conference on Information Technology Research (ICITR)
[9] Oro Ermelinda, Massimo Ruffolo and Mostafa Sheikhalishahi 2018 Language Identification of Similar Languages using Recurrent Neural Networks ICAART (2) pp 635-640
[10] https://www.kaggle.com/datasets/basilb2s/language-detection?datasetId=1150837