Mapana – Journal of Sciences
2024, Vol. 23, No. 4, 121-136
ISSN 0975-3303 | https://doi.org/10.12723/mjs.71.7

Fake News Detection in Low Resource Language Using Machine Learning Techniques and SMOTE

Rajalakshmi Sivanaiah*, Angel Deborah Suseelan*, Sushanth Dilli Baskar* and Swathika Durairaj*

Abstract

Fake content dissemination is a significant challenge in the era of digital information. This paper discusses the critical issues in detecting fake content in news articles of low-resource languages, specifically focusing on the Tamil language, where the availability of labeled data and advanced natural language processing tools are limited. We employ traditional machine learning models to mitigate this problem, with particular emphasis on detecting and classifying fake and real content in the context of Tamil news. Our study explores the performance of different models like logistic regression (F1 score: 91%), support vector machines (SVM) (F1 score: 91%), naive Bayes (F1 score: 89%), k-nearest neighbors (KNN) (F1 score: 70%), decision trees (F1 score: 91%), random forests (F1 score: 86%) and passive-aggressive classifier (F1 score: 89%). By conducting a comprehensive comparative analysis of these models within the challenging linguistic environment of Tamil, we aim to provide insights into their suitability for detecting fake content in low-resource languages and draw meaningful comparisons between their performance.

* Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai – 603110, Tamil Nadu, India; rajalakshmis@ssn.edu.in, angeldeborahs@ssn.edu.in, sushanth2110209@ssn.edu.in, swathika2110791@ssn.edu.in

Keywords: Fake news classification · Low resource languages · Machine learning models · Comparative analysis.

1. Introduction

In an age characterized by the relentless surge of digital information, the rampant dissemination of fake content in images, audio, videos, and text has evolved as a formidable challenge, posing severe threats to society worldwide. The impact of fake news is ubiquitous and insidious, penetrating even the linguistic boundaries of low-resource languages, where resources and tools for effective detection are markedly scarce. This research paper delves into this pressing concern, specifically concentrating on the vital issues of detecting fake content in the context of less resource languages, with a particular focus on the Tamil language.

Low resource languages, as characterized by their deficiency in linguistic re-sources and their underdeveloped Natural Language Processing (NLP) tools, confront unique and pressing challenges in the realm of fake news detection. These languages, in stark contrast to their high-resource counterparts like English, often struggle to access the required infrastructure and resources to combat the menace of fake news effectively. It is in this challenging linguistic landscape that our research paper aims to make significant contributions. Tamil is a very old language that contains a rich heritage and culture and serves as a compelling focal point for our study. Boasting over 70 million speakers worldwide, Tamil holds immense cultural and historical significance, particularly in Tamil Nadu, a state in India and the island nation of Sri Lanka. Nevertheless, despite its widespread influence and cultural importance, Tamil remains classified as a low-resource language in the context of NLP and fake content detection.

This research paper seeks to bridge this gap by assessing the need and performance of traditional machine-learning techniques in the context of low-resource languages [7]. We employ a range of machine learning models to achieve this, like k-nearest neighbors (KNN), logistic regression, naive Bayes, support vector machines (SVM), decision trees, and random forests, which are well-established and widely utilized in various classification tasks. The performance of these models is evaluated using the F1 score, a metric that balances

precision and recall, providing a robust measure of their effectiveness in identifying fake news [4].

The central objective of our research is to conduct a comprehensive comparative analysis of these models to determine their suitability for detecting fake news in low-resource languages. By doing so, we aim to contribute valuable insights into the development of effectual strategies for combating misinformation in linguistic environments with limited resources. We have discussed the related work, methodologies used, results, and discussions arising from our analysis of the machine learning models in the context of low-resource languages in the following sections. The ultimate goal is to advance our collective understanding of fake news detection and mitigation in less resource languages.

2. Related Work

In the digital realm, the scarcity of the tools used to identify fake content in low-resource languages heightens the risk of rampant disinformation, eroding online information reliability and digital communication trustworthiness. The imbalance in global linguistic resource availability severely hinders most languages, termed "low-resource languages," from accessing essential Natural Language Processing (NLP) tools. To address this, Lin et. al, conducted a study in 2020, focusing on creating NLP tools for low-resource languages, incorporating techniques such as machine translation, part-of-speech tagging, and lexicon development, sometimes with crowdsourced assistance. Their research introduces a comprehensive framework that streamlines language resource creation for low-resource languages, mitigating the time and cost constraints inherent in starting from scratch [1]. Sumeet Dua and Xian Du emphasized the pivotal role of Machine Learning (ML) algorithms in enhancing software accuracy without requiring direct reprogramming [2]. In our research, we employ six algorithms [3], including Decision Trees, for fake news classification. Decision Trees provide a structured approach, enabling the identification of crucial variables, visualization of variable relationships, and the creation of new features for accurate target variable prediction, aligning with the objectives of our study.

Our goal is to identify fake news, with a specific focus on Tamil, an Indic language. Our approach places a significant emphasis on the textual content of messages, along with user and network-related features. To process text effectively, we employ natural language processing tools and techniques [5]. The initial model training is conducted with English-language data, which is subsequently translated into Tamil [6]. The model is then fine-tuned using these translated datasets and utilized in machine learning models to assess the credibility of news articles. To facilitate further analysis, all data must be converted into numerical form. For this purpose, we utilize machine learning models, which are particularly effective in handling text-based information. We have also studied the application of deep learning techniques and identified that those techniques perform well in large datasets.

3. Methodology

3.1 Tamil News Dataset

The main objective of this work is to achieve high accuracy in the detection of fake news from real news, especially in low-resource languages [10]. In this context, the Tamil language was chosen for the study. The dataset used in this paper comprises a total of 14,564 rows, with each row containing news headlines in four features, including English text, Tamil text, Index, and IsFake [8]. To streamline the dataset for this study, the 'English' and 'Index' columns, which do not hold any substantive significance, are removed. As we have limited data, we have experimented with machine learning methods.

The resulting dataset consists of 11,662 instances of real news and 2,902 instances of fake news. As evident from the data sample distribution, the dataset is not balanced. To overcome this, oversampling is performed on the fake news class to balance the number of fake and real articles that left us with 23324 rows of data [9]. Oversampling represents a data preprocessing method designed to rectify imbalanced class distributions within a dataset [19]. Its primary objective is to address scenarios where one or more classes are underrepresented compared to others. In our specific dataset, there exists a notable skew toward real news instances, with a ratio of approximately 4:1 in favor of real news over fake news. To mitigate this imbalance, we

employ the Synthetic Minority Over-sampling Technique (SMOTE) technique, which involves the generation of synthetic data points for the minority class by interpolating between existing instances. This approach ensures a more equitable representation of the minority class, ultimately enhancing the capacity of the model to learn both real and fake news samples. Subsequently, the dataset was split in the ratio of 80:20 samples for training and testing purposes. This partitioning allows for the development and evaluation of a robust detection model for fake news in the Tamil language. Also, another dataset was collected to find and remove the stopwords in Tamil, which contains 125 Tamil stopwords. Figure 1 shows the sample data for real news and Figure 2 shows the sample data for fake news in Tamil language. The sample of stop words in Tamil is shown in Figure 3.

Index	Tamil text	Is Fake
0	சபரிமலையில் குவியும் பக்தர்கள்: ஐயப்பனை இரவு 11 மணிவரை தரிசிக்க அனுமதி	0
1	ஆரியங்காவு தர்மசாஸ்தா அன்னை புஷ்கலா தேவி திருக்கல்யாணம் கோலாகலம்	0
2	புலிவாகனத்தில் மகரசங்கராந்தி பிரவேசம் - யாருக்கு ராஜயோகம்... யாருக்கு நஷ்டம் தெரியுமா	0
3	தன்வந்திரி பீடத்தில் அனுமன் ஜெயந்தி: நாமக்கல் ஆஞ்சநேயருக்கு அணிவிக்க 100008 வடை மாலை தயார்	0
4	தடுப்பூசி போடலியா.. பேங்க், ஹோட்டல் கூட போக முடியாது.. தடை போட்ட ஹரியானா அரசு	0
5	ஹரியானா நிலச்சரிவில் சிக்கி 4 பேர் பலி.. மற்றவர்கள் நிலை என்ன?	0
6	"முன்னெச்சரிக்கையே இல்லை".. சாடிய முதல்வர் ஸ்டாலின்.. "ரொம்ப கஷ்டம்".. வானிலை மையம் தந்த பதில்!	0
7	சைதாப்பேட்டை அரசு பயிற்சி மையத்தில் 34 பேருக்கு கொரோனா.. மாணவர்கள் உடல்நிலை எப்படி உள்ளது?	0
8	Happy New Year 2022: உலகிலேயே முதலில் புத்தாண்டை வரவேற்கும் நாடு எது தெரியுமா? கடைசி நாடு இதுதான்	0
9	இந்தியாவை மிக உயர்ந்த இடத்துக்கு கொண்டு செல்ல வேண்டும்.. ஆளுநர் தமிழிசை வேண்டுகோள்	0
10	'ஓமிக்ரான் வைரஸ்: 3ஆம் அலை உறுதி, உடனடி நடவடிக்கை தேவை..' முதல்வருக்கு ரவிக்குமார் எம்பி பரபர கடிதம்	0

Fig. 1. Dataset for True news

Index	Tamil text	IsFake
11806	ஜார்ஜ் வாஷிங்டன் மேற்கோள் காட்டுகிறார், "எந்தவொரு நாடும் தனது குடிமக்களை துப்பாக்கிகள் மூலம் அவநம்பிக்கை கொள்ளும்போது ... அது இனி அதன் குடிமக்களை நம்பாது, ஏனெனில் அத்தகைய அரசாங்கம் தீய திட்டங்களைக் கொண்டுள்ளது."	1
11807	"கடந்த நூற்றாண்டில் ஹிட்லர், ஸ்டாலின், மாவோ போன்றோர் இந்த நூற்றாண்டில் ஒரு அழிவுகரமான நபர் டிரம்ப். அவர்களை விட பல மில்லியன் இறப்புகளுக்கு அவர் காரணமாக இருக்கலாம்."	1
11808	சீனா மீதான அமெரிக்க வரிகள் அமெரிக்காவில் "யாரையும் காயப்படுத்தவில்லை".	1
11809	பிலடெல்பியா துப்பாக்கிச் சூடு "போலி" என்றும் காவல்துறை அதிகாரிகள் "போலி ரத்தத்தை தங்கள் மீது தெளித்துக் கொண்டனர்" என்றும் கூறுகிறார்.	1
11810	அடுத்த நாள் சமைக்கப் பயன்படுத்தப்படும் முன் வெட்டப்பட்ட வெங்காயம் "ஒரே இரவில் கூட மிகவும் நச்சுத்தன்மையுடையதாக மாறும் மற்றும் நச்சு பாக்டீரியாவை உருவாக்குகிறது, இது அதிகப்படியான பித்த சுரப்பு மற்றும் உணவு நச்சுத்தன்மையின் காரணமாக பாதகமான வயிற்று நோய்த்தொற்றுகளை ஏற்படுத்தக்கூடும்."	1
11811	சிஎன்என் தலைவர் ஜெஃப் ஜுக்கர், ஜெஃப்ரி எப்ஸ்டீனுடனான பில் கிளிண்டனின் உறவைக் குறைத்து மதிப்பிடுமாறும், எப்ஸ்டீன் "இவை அனைத்திலும் டிரம்பின் பங்குதாரர்" போல் தோற்றமளிக்குமாறும் ஊழியர்களுக்கு உத்தரவிட்டார்.	1
11812	சான் ஃபிரான்சிஸ்கோவின் வீடற்ற மக்களில் "பெரும்பாலானோர்" டெக்சாஸிலிருந்து வந்தவர்கள் - இது எங்களுக்குத் தெரியும். வெறும் (ஒரு) சுவாரஸ்யமான உண்மை."	1
11813	ஜெஃப்ரி எப்ஸ்டீன் இறக்கவில்லை என்று சொல்லுங்கள்.	1
11814	மேற்கோள்கள் ரெப். கெவின் மெக்கார்த்தி "ஜப்பானில் வீடியோ கேம்கள் இல்லாததால் அங்கு வெகுஜன துப்பாக்கிச் சூடு சம்பவங்கள் இல்லை."	1
11815	ஜனாதிபதி டொனால்ட் டிரம்ப் "லத்தீன் இனத்தை அழிப்பதைப் பற்றி பேசுகிறார்" என்று கூறுகிறார்.	1

Fig. 2. Dataset for Fake news

ஒரு
என்று
மற்றும்
இந்த
இது
என்ற
கொண்டு
என்பது
பல
ஆகும்
அல்லது
அவர்
நான்
உள்ள
அந்த

Fig. 3. Stopwords in Tamil

3.2 Model Architecture

In this research, we have employed a comprehensive approach utilizing seven distinct machine-learning techniques. This diverse array of methods enhances the robustness of our findings. Specifically, we have harnessed seven different machine learning models, which encompass Logistic Regression, Passive-Aggressive Classifier, Support Vector Machine, Decision Tree, K-Nearest Neighbor, Random Forest, and Multinomial Naive Bayes [11,12]. The output layer identifies the likelihood of the input sentence being fake or real news. Figure 4 shows the architecture of the system used to detect the fakeness in the news dataset.

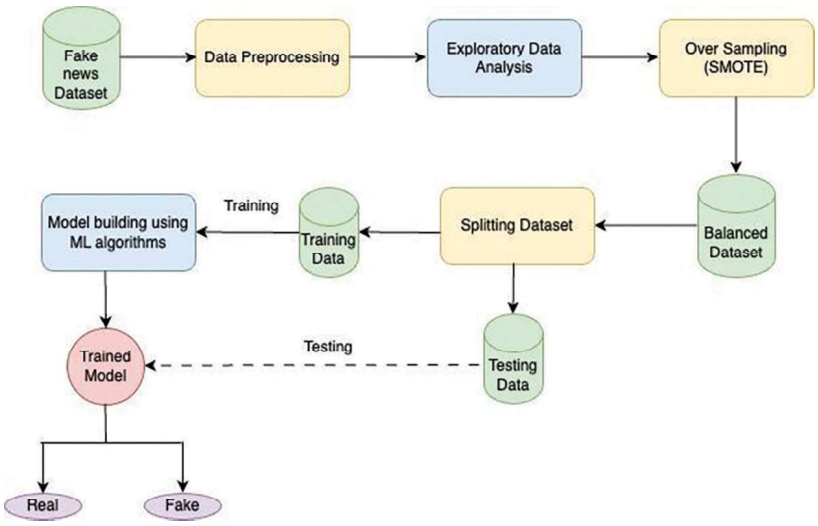

Fig. 4. Model Architecture

The input dataset is preprocessed to remove the stopwords and unwanted characters. The data is explored and analyzed for the variations in the dataset. It is observed that the data set is not balanced. The data is distributed with more real news samples when compared to the fake news samples. We have used the data augmentation technique by oversampling the data using SMOTE to balance the dataset. Then the balanced dataset is split into training and testing phases. We have built machine learning (ML) models using different algorithms with the data samples in training. The trained model is used with the test data to predict the output labels (fake/real). We have observed that deep learning models are not performing well in this case as the dataset is small in size.

The various machine-learning techniques used for training and learning the data as listed as follows:

- Logistic Regression: Logistic Regression is a strong performer with high accuracy. It's effective for binary classification tasks, and its simplicity makes it interpretable and easy to implement. However, it may not handle highly non-linear relationships in the data as effectively as other models.
- SVM (Support Vector Machine): SVM performs well with high accuracy [14]. It is suitable for complex classification tasks and can capture intricate decision boundaries. It is computationally intensive for large datasets.
- Naive Bayes: It is a simple and efficient probabilistic classifier. However it assumes independence between features, which may not hold in all real-world scenarios [13].
- KNN (K-Nearest Neighbors): K-Nearest Neighbors (KNN) is a fundamental machine learning algorithm used for both classification and regression tasks [15]. It is a non-parametric and instance-based learning method. It does not make any underlying assumptions about the data distribution. Instead, KNN makes predictions by considering the " k" nearest data points to the target point in the feature space.
- Decision Tree: It can capture the complex relationships that occur in the data and provide interpretability. It is prone to overfitting with deep trees.
- Random Forest: This combines multiple decision trees to reduce overfitting and thereby improves robustness, but it is computationally expensive.
- Passive Aggressive: This classifier is a machine learning algorithm known for its effectiveness in online learning scenarios. It considers the data as it arrives incrementally.

4. Implementation

The proposed model algorithms take its input as a CSV (Comma comma-separated values) file, initially subjecting the data to preprocessing and cleaning such as removing stopwords, punctuation, tokenization, removing small words, etc. Recognizing the imbalance

of samples in the dataset, oversampling is conducted to ensure a more balanced representation of real news. Figures 5 and 6 show the distribution of the data samples before and after applying the data augmentation technique. Subsequently, the dataset is partitioned into training and testing sets. To facilitate machine learning model comprehension, we apply the CountVectorizer() method, which transforms the textual data into numerical vectors.

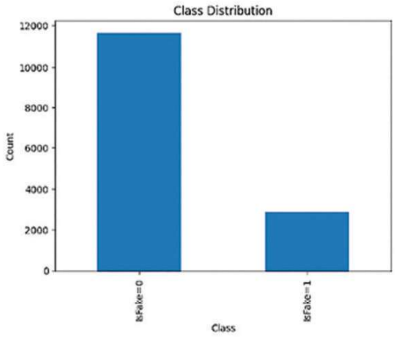

Fig. 5: Before Data Augmentation

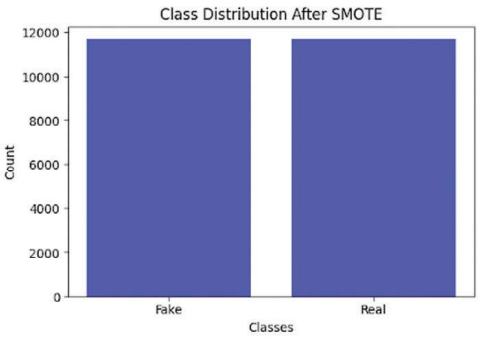

Fig.6. After Data Augmentation

Once the data is prepared, it is fed into the classifiers defined within a pipeline. Each classifier processes the training dataset, and accuracy scores are predicted, contributing to a comprehensive evaluation of the model's performance [16]. This multifaceted approach enables us to assess the effectiveness of different machine learning techniques in detecting fake content in Tamil language news.

5. Result and Analysis

The performance evaluation metrics of various machine learning models in the context of fake Tamil news detection are shown in Table 1.

The evaluation metrics include Accuracy, Precision, Recall, and F1 Score, which collectively offer the details of the models' capabilities in correctly classifying instances and F1 score is used for balancing the tradeoff between Precision and Recall. Logistic Regression, SVM, and Random Forest exhibit high accuracy and F1 score making them strong performers in this classification task. These models are effective at correctly classifying both real and fake data.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	91	89	93	91
SVM	91	90	92	91
Naive Bayes	89	90	87	89
KNN	57	54	68	70
Random Forest	91	90	92	91
Decision Tree	85	82	90	86
Passive Aggressive	88	85	91	89

Table 1. Comparison of evaluation metrics

Naive Bayes demonstrates good precision. This probabilistic classifier is suitable for the task but slightly less accurate compared to Logistic Regression, SV, M, and Random Forest. KNN has notably lower accuracy (57%) compared to the other models. The precision is also relatively lower (54%). KNN's poor performance in this context may be attributed to its sensitivity to noise and an inappropriate choice of "k". We have taken the "k" value as 10. Random Forest and Decision Tree models perform reasonably well with accuracy scores of 91% and 85%, respectively. These models capture complex and internal relationships in the data, with Random Forest outperforming Decision Tree. The Passive-aggressive classifier obtains an accuracy of 88% with good precision and F1 score. It performs well but is slightly less accurate compared to Logistic Regression and SVM and also has less F1 score.

The confusion matrix is a crucial method for assessing the model's performance on the predictions when tested on a dataset. It provides a clear picture of the model's performance by detailing true positives, true negatives, false positives, and false negatives [17,18]. The provided figures from 7 to 13 display the confusion matrix plots for several classifiers, including Logistic Regression, Decision Tree, Naive Bayes, KNN, SVM, Random Forest, and Passive Aggressive. It's worth noting that Logistic Regression achieved the highest sum of true positives and true negatives (4258), indicating its strong predictive accuracy. Support Vector Machine follows closely with a sum of 4254. KNN on the other side shows the lowest sum of true positives and true negatives (2666) among the models, suggesting

that it may not be as accurate in making correct predictions as the other classifiers.

Figure 7 describes the confusion matrix of Logistic regression (LogR). The confusion matrix reveals a count of 2086 true negative instances, 2172 true positive instances, 255 false positive instances, and 152 false negative instances. Figure 8 shows the confusion matrix of SVM. The confusion matrix reveals a count of 2105 true negative instances, 2149 true positive instances, 236 false positive instances, and 175 false negative instances. Figure 9 depicts the confusion matrix of Naive Bayes (NB). The confusion matrix reveals a count of 2124 true negative instances, 2019 true positive instances, 217 false positive instances, and 305 false negative instances. Figure 10 gives the confusion matrix of KNN. The confusion matrix reveals a count of 345 true negative instances, 2321 true positive instances, 1996 false positive instances, and 3 false negative instances.

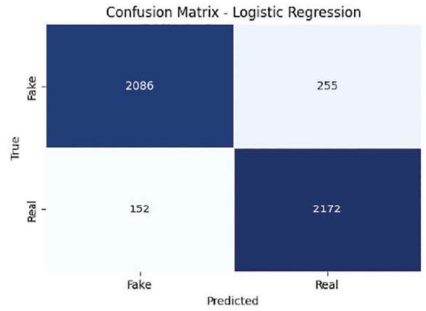

Fig. 7. Confusion matrix for LogR

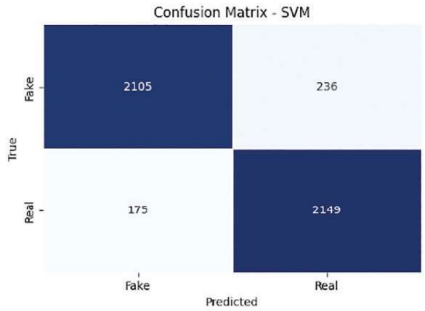

Fig. 8. Confusion matrix for SVM

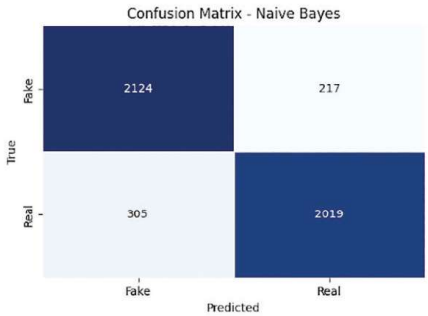

Fig. 9. Confusion matrix for NB

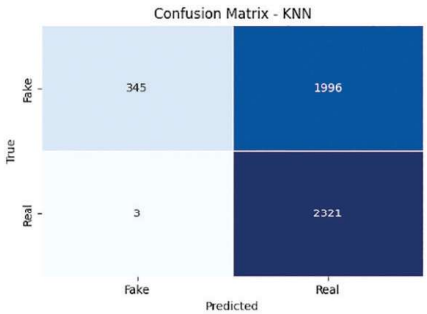

Fig. 10. Confusion matrix for KNN

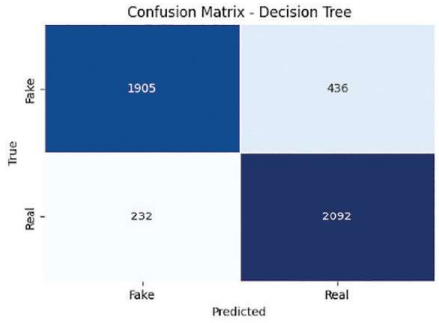

Fig. 11. Confusion matrix for DT

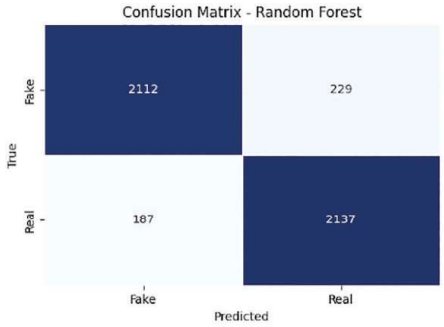

Fig. 12. Confusion matrix for RF

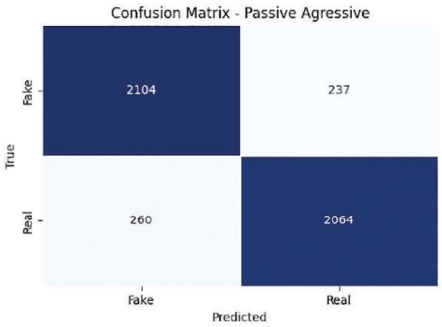

Fig. 13. Confusion matrix for PA

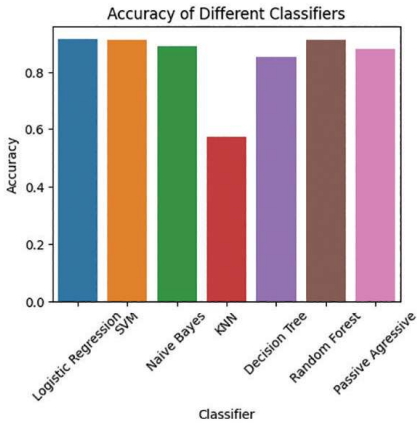

Fig. 14. Comparison of Accuracy with different classifiers

The confusion matrix of the decision tree (DT) is shown in Figure 11. This matrix reveals a count of 1905 true negative instances, 2092 true positive instances, 436 false positive instances, and 232 false negative instances. Figure 12 describes the confusion matrix of random forest (RF). The confusion matrix reveals a count of 2112 true negative instances, 2137 true positive instances, 229 false positive instances, and 187 false negative instances. The confusion matrix of the passive-aggressive (PA) classifier is given in Figure 13. The confusion matrix reveals a count of 2104 true negative instances, 2064 true positive instances, 237 false positive instances, and 260 false negative instances. Figure 14 shows the graphical comparison of the accuracy metric for all the classifiers.

6. Conclusion

In conclusion, this research contributes to the usage of ML models for combating misinformation in linguistic environments with limited resources, offering valuable guidance to researchers and practitioners in this field. Specifically, we have delved into the critical issue of fake content detection in low-resource languages, with the Tamil language as our focal point in our maiden attempt. In this challenging linguistic environment, where labeled data and advanced natural language processing tools are scarce, we embarked on the exploration of traditional machine learning models. Our study has unveiled significant findings, with logistic regression demonstrating a noteworthy F1 score of 91%, closely followed by support vector machines (SVM) at 91%. This comparative examination has also shed light on the limitations of KNN, which yielded a lower F1 score of 70%.

Meanwhile, Naive Bayes, Decision Tree, and Random Forest exhibited competitive F1 scores of 89%, 91%, and 86%, respectively, while the Passive Aggressive Classifier achieved an F1 score of 89%. These results describe invaluable insights into machine learning models' performance for detecting fake content in low-resource languages, offering guidance to researchers and practitioners alike. As the battle against misinformation intensifies, our research contributes to the development of effective strategies tailored to linguistic environments with limited resources. Ultimately, our goal is to bolster the defenses against the dissemination of fake content, safeguarding the integrity of information ecosystems. This also promotes informed decision-making in diverse linguistic contexts.

Funding information:

We have not received funding from any organization.

Author contributions:

Authors 1 and 2 contributed to idea formulation, literature survey, design, testing, and prepared the main manuscript text. Authors 3 and 4 have done the literature survey, implementation, testing, and prepared the results. All authors reviewed the manuscript.

Conflict of Interest:

The authors have declared no conflict of interest.

Acknowledgement:

We would like to thank the Department of Computer Science and Engineering and the management of Sri Sivasubramaniya Nadar College of Engineering for providing the facilities to do this research work.

References

[1]. D. Lin, Y. Murakami, and T. Ishida, "Towards Language Service Creation and Customization for Low-Resource Languages," *Information*, vol. 11, no. 2, p. 67, 2020. doi: 10.3390/info11020067.

[2]. S. Dua and X. Du, *Data Mining and Machine Learning in Cybersecurity*, New York: Auerbach Publications, 2016. doi: 10.1201/b10867.

[3]. Z. Khanam, B. N. Alwasel, H. Sirafi, and M. Rashid, "Fake news detection using machine learning approaches," in *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012040, IOP Publishing, 2021. doi: 10.1088/1757-899X/1099/1/012040.

[4]. Q. Nan, J. Cao, Y. Zhu, Y. Wang, and J. Li, "MDFEND: Multi-domain fake news detection," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 3343-3347, 2021. doi: 10.1145/3459637.3482139.

[5]. A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," arXiv preprint arXiv:2006.07264, 2020. doi: 10.48550/arXiv.2006.07264.

[6]. D. Kakwani, A. Kunchukuttan, S. Golla, N. C. Gokul, A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4948-4961, 2020. doi: 10.18653/v1/2020.findings-emnlp.445.

[7]. S. Gokila, S. Rajeswari, and S. Deepa, "TAMIL-NLP: Roles and impact of machine learning and deep learning with natural language processing for Tamil," in *2023 Eighth International Conference on Science Technology Engineering and*

Mathematics (ICONSTEM), pp. 1-9, IEEE, 2023. doi: 10.1109/ICONSTEM56934.2023.10142680.

[8]. AjhayAk, "Dataset: FakeNewsDetectionTamil," 2023. [Online]. Available: https://github.com/AjhayAk/FakeNewsDetectionTamil.

[9]. R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, pp. 243-248, 2020. doi: 10.1109/ICIC S49469.2020.239556.

[10]. M. Fayaz, A. Khan, M. Bilal, and S. U. Khan, "Machine learning for fake news classification with optimal feature selection," *Soft Computing*, vol. 26, no. 16, pp. 7763-7771, 2022. doi: 10.1007/s00500-022-06773-x.

[11]. S. Gupta and P. Meel, "Fake news detection using passive-aggressive classifier," in *Inventive Communication and Computational Technologies: Proceedings of ICICCT 2020*, Lecture Notes in Networks and Systems, vol. 145, pp. 155-164, Springer, Singapore, 2021. doi: 10.1007/978-981-15-7345-3_13.

[12]. M. J. Awan et al., "Fake news data exploration and analytics," *Electronics*, vol. 10, no. 19, p. 2326, 2021. doi: 10.3390/electronics10192326.

[13]. M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pp. 900-903, IEEE, 2017. doi: 10.1109/UKRCON.2017.8100379.

[14]. M. G. Hussain, M. R. Hasan, M. Rahman, J. Protim, and S. A. Hasan, "Detection of Bangla fake news using MNB and SVM classifier," in *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pp. 81-85, IEEE, 2020. doi: 10.1109/iCCECE49321.2020.9231167.

[15]. T. Mladenova and I. Valova, "Analysis of the KNN classifier distance metrics for Bulgarian fake news detection," in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1-4, IEEE, 2021. doi: 10.1109/HORA52670.2021.9461333.

[16]. N. Smitha and R. Bharath, "Performance comparison of machine learning classifiers for fake news detection," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 696-700, IEEE, 2020. doi: 10.1109/ICIRCA48905.2020.9183072.

[17]. D. Mangal and D. K. Sharma, "A Framework for Detection and Validation of Fake News via Authorize Source Matching," in *Micro-Electronics and Telecommunication Engineering: Proceedings of 4th ICMETE 2020*, Lecture Notes in Networks and Systems, vol. 179, Springer, Singapore, 2021. doi: 10.1007/978-981-33-4687-1_54.

[18]. S. R. Indarapu et al., "Comparative analysis of machine learning algorithms to detect fake news," in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, pp. 591-594, IEEE, 2021. doi: 10.1109/ICSPC51351.2021.94 51690.

[19]. R. Sivanaiah et al., "Fake News Detection in Low-Resource Languages," in M. A. K. et al., *Speech and Language Technologies for Low-Resource Languages*, SPELLL 2022, Communications in Computer and Information Science, vol. 1802. Springer, Cham. doi: 10.1007/978-3-031-33231-9_23.