Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine Learning Techniques for Effective Multilingual Text Classification

[1]Pranav Rajendra Patil, [2]Dr. Monali Y. Khachane
[1]Research Scholar
KCES's M. J. College (Autonomous),
Jalgaon, Maharashtra, India
[2]Asst. Professor
Dept. of Computer Science,
(Dr. Annasaheb G.D. Bendale Mahila Mahavidhyalaya, Jalgaon, Maharashtra,India)

ARTICLE INFO

Received: 10 Feb 2025

Revised: 15 Mar 2025

Accepted: 22 Mar 2025

ABSTRACT

The rapid growth of the internet and digital communication has led to an unprecedented increase in textual content across numerous languages worldwide. As a result, the field of Natural Language Processing (NLP) faces a critical challenge in accurately identifying and recognizing languages, especially in multilingual environments. This study highlights the critical role of multilingual identification and recognition systems in enabling communication across diverse linguistic environments. These systems serve as foundational tools for a variety of applications, such as translation services and speech recognition, which rely on accurately identifying and understanding languages spoken or written in various contexts. This paper presents a systematic evaluation of such systems, focusing specifically on English, Hindi and Marathi languages. A range of approaches, including machine learning models, deep learning, and NLP, were analyzed. The inclusion criteria for this study consisted of research publications from the years 2019-2024. Overall, the findings underscore the importance of identification of multilingual languages. This review offers substantial insights for both researchers and practitioners engaged in the advancement of robust multilingual identification and recognition systems designed for English, Hindi and Marathi languages. Further study is imperative to overcome current obstacles and augment the efficiency of these multilingual identification and recognition techniques.

Keywords: Multilingual identification, Deep learning, Machine learning.

INTRODUCTION

Identifying languages is a major challenge in Natural Language Processing (NLP). With the internet growing, more and more content is being created in languages other than English [1]. For tasks like searching and organizing text automatically, it's important to identify the language in real-time [2]. In multilingual countries like India, where people often speak more than one language, recognizing languages is especially important. While many languages are spoken in India, the constitution officially recognizes 22 of them [3, 4].

A bilingual or multilingual community is formed when individuals use more than one language as a medium of communication. This can happen when people converse in their native language or any other natural language that has official, national, or international significance. In daily online conversations with friends and family, multilingual people frequently favor mixed-language

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

constructions [5]. The use of code-mixing, code-switching, or a blend of the two to productively communicate is the defining feature of the text corpus based on social media [6-9].

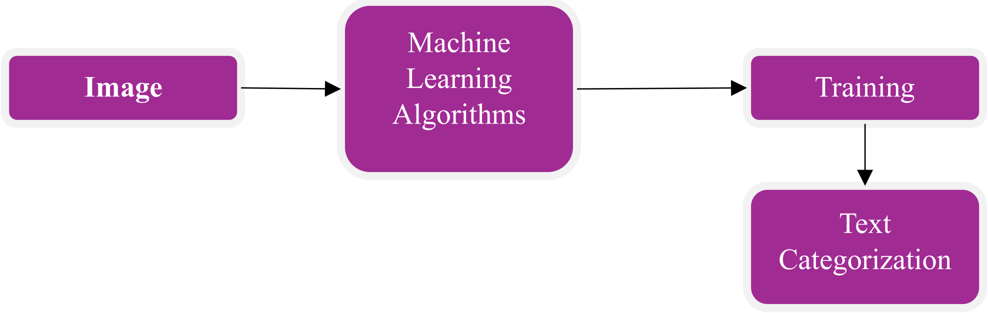


Figure 1. Text Classification

1.1 Automatic Language Identification

Reliable language identification is becoming more important for several reasons, such as the rising popularity of hands-free, voice-operated devices for human interaction and the necessity to support the coexistence of diverse languages in a more interconnected world [12, 13].The goal of automatic language identification is to detect the language or languages used in text texts without human intervention. Documents can have one or more languages [14]. A framework for text document language identification is shown in Figure 1.

1.2 Multilingual handwritten text recognition (MLHTR)

Multilingual communication, data entry activities, and the digitization of historical records are just a few of the areas where this technology is vital [16]. For example, India is home to a plethora of multilingual publications due to the country's dozens of widely spoken languages [17]. Many real-world contexts including healthcare, schools, insurance, banking, the government and the financial sector present the challenge [18].

1.3 Applications of Language Identification

The classification of text has been employed for several objectives, including improving the comprehension of textual materials, organizing textual documents for information retrieval, and categorizing notes and emails, among other assignments. The identification of multilingual documents serves various purposes, including filtering data to improve the quality of input data and extracting linguistic information from the web and publications. The process of identifying multilingual documents can also serve to extract bilingual texts from internet sources [28-29].

1.4 Challenges Involved in Language Identification

There are two types of text available for the language identification task: multilingual and monolingual. Processing monolingual texts is remarkably straightforward, as it only demands proficiency in a single language. Processing multilingual documents, however, is more difficult since it requires an understanding of several languages and the challenges associated with their interdependency. The following are the primary difficulties encountered in language identification [30]. Figure 2 shows the various challenges related to language identification.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

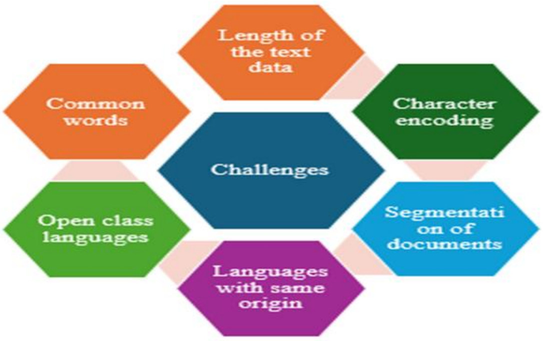


Figure 2: Challenges of Language identification [30]

- **Noisy text:** The use of short forms of words and tags, known as abbreviations, is regarded as unnecessary information in the text. Users sometimes use abbreviations or codes when writing content due to space limitations and the inconvenience of typing, resulting in a text that can only be understood by its writer. Managing such a noisy dataset presents a really challenging task [31-32].
- **Length of the text data:** The document's text data must be sufficiently large to enable the efficient use of n-gram algorithms to determine the language with accuracy. However, the messages in social media programs such as WhatsApp and Twitter are limited in size, making them tough to deal with.
- **Character encoding:** Character sets are used differently in different natural languages. Character encoding, sometimes referred to as character set, is a technique that represents characters by combining several symbols. There are several encoding schemes available, including Unicode and ASCII. There can be more than one language encoding for a single text document. Managing encoding discrepancies can be difficult and calls for a flexible method [33].
- **Segmentation of documents:** When text is written in multiple languages and used interchangeably, the document is said to be multilingual. The partition of text into several languages, which is required to separate the contents of the document, is the main issue in this work. Once the segments have been identified, the content can be collected for further processing.
- **Common words:** When two language communities interact culturally, one language acquires vocabulary from the other. This process is known as linguistic borrowing. These words are called borrowings or loanwords.
- **Languages with the same origin:** Linguistically related languages commonly exhibit shared writing systems and grammatical characteristics because of their shared ancestry. Both Hindi and Marathi employ the Devanagari script. Distinguishing between these languages poses a substantial difficulty in language identification tasks.

1.5 Need for Multilingual Text Recognition system for Indian languages (English, Hindi, Marathi)

Text recognition has made significant advancements in recent years, due to the rapid development of deep learning. With the extensive use of English, current research predominantly concentrates on the recognition of English text. Existing recognition algorithms [34-41] were primarily developed for English texts. However, existing research mainly concentrates on writings written in commonly used scripts, such as English or Chinese, while considering multilingual literature. Practical scene images frequently include texts from various scripts rather than being limited to just one script. Multilingual text recognition is more suited for practical application settings since it can simultaneously recognize texts in multiple scripts [42].

The process of identifying and separating connected characters in Devanagari scripts is accomplished using fuzzy multi-factorial analysis. Many character segmentation techniques are unsuccessful in segmenting touching characters, sometimes incorrectly identifying them as single characters. There is

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

currently no universal character segmentation approach that can be applied to all Indian languages [45].

1. Review Methodology

A detailed review of studies on this topic was done using the SCOPUS database, covering works published between 2019 and 2024. This review focused on records categorized as articles, journals, and publications in Scopus. Table 1 lists the keywords used to search for relevant papers in the database.

Table 1: Searching Keywords

Databases	**Keyword used**
Scopus	TITLE-ABS-KEY ("Text classification" OR "Language Identification" OR "Multilingual Identification") AND TITLE-ABS-KEY ("Machine learning" OR "Natural language processing") AND TITLE-ABS-KEY ("English" OR "Hindi" OR "Marathi")

Source: Authors own elaboration

For the literature analysis, only the records marked as articles are further marked countable and evaluated for assessment. The below-presented table 2 exclusively examined records that met the specified inclusion and exclusion criteria:

Table 2: The criteria for determining what is Included and Excluded

Criterion	**Inclusion**	**Exclusion**
Keywords	Records conferring the relationship between text classification and Multi identification, Machine learning and Natural language processing, and English, Hindi and Marathi.	Records excluded in which variables have no relation
Doc Type	Conf. paper, article, conf. review, Review paper	Book series, book, chapter in book
Language	English	Other than English
Timeframe	Concerning 2019-2024	<2019
Category	Open Access	Paid Access

Source: Authors own elaboration

LITERATURE ANALYSIS

The evaluation contributes significantly to the understanding of ML, DL and NLP applications in multilingual contexts. It sheds light on the strengths and limitations of these systems, aiding in the refinement of recognition and identification capabilities for diverse linguistic environments. Additionally, this analysis serves as a foundation for ongoing innovation and development in multilingual identification and recognition systems leveraging ML, DL and NLP techniques. It provides valuable insights for researchers and practitioners seeking to enhance the performance of such systems across different languages and cultural contexts.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

(a) Identification through Machine Learning

In recent years, researchers have actively explored various ML approaches for text classification tasks. **Najahan Binti Mohd Rashidi et al. (2024) [46]** explored machine-learning approaches and found that the CountVectorizer method outperformed TF-IDF in their dataset. Similarly**, Gholami et al. (2023) [47]** focused on leveraging graph neural network (GNN) models to effectively capture topological information in text data.

Salh et al. (2023) [48] examined the efficiency of machine learning methods, particularly Support Vector Machine and Random Forest, in detecting fake news, especially in low-resource languages like Kurdish. Similarly, **Alfartosy et al. (2023)[49]**demonstrated strong performance with their ML approach, surpassing previous methodologies, and **Das et al. (2023) [50]** made significant contributions to sentiment analysis, offering valuable insights for future research. **Al-onazi et al. (2023) [51]** attained exceptional results by combining ML and deep learning techniques in their system analysis task.

To et al. (2020) [59] and **Phat et al. (2020) [60]** reported promising outcomes in their respective studies, contributing to the evolving landscape of ML-based text classification. Additionally, **Sarwar et al. (2020) [61]** highlighted the robustness of the K-nearest neighbours (KNN) algorithm across languages.

Key findings from the survey of material available in the open literature

Table 3. shows the comparative study between several techniques applied for identification.

Author	Technique	Accuracy	Outcomes
Najahan Binti Mohd Rashidi et al., (2024) [46]	SVM	92%	The result suggested that CountVectorizer performs better for the dataset.
Gholami et al., (2023) [47]	Graph neural network	86.36%	The results illustrated how the application of GNN models contributes to achieving high scores in text classification by effectively capturing the topological information between textual data.
Salhet al., (2023) [48]	SVM, RF	91%	The results indicated that machine learning techniques can accurately identify false information in languages with limited resources, such as Kurdish, even in challenging circumstances.
Alfartosy et al., (2023) [49]	Logistic Regression	98%	The results revealed strong performance, with the proposed method surpassing recent approaches.
Das et al., (2023) [50]	Long-Short Term Memory, Bidirectional LSTM	86.43%	The study's findings significantly advance the field of sentiment analysis by providing insightful information for further investigation and useful applications.
Wasim et al., (2023) [52]	RNN,SVM	90%	The ensemble-based approach demonstrates high performance on two benchmark datasets, BET and UFN.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Fkih et al., (2023) [53]	Neural network	88%	The results obtained indicate that the Neural Network outperforms other models and delivers strong performance when utilizing hybrid features.
Polatet al., (2022) [57]	SVM,RF,RNN	82%	The experimental results indicate that the demographics of deputies can be effectively estimated using NLP, ML, and Deep Learning approaches.
Barua et al., (2021) [58]	LR, SVC, DT	98.13%	The result revealed that the Support Vector Classifier (SVC) using unigram + bigram + trigram feature space obtained the highest weighted f1-score of 97.60%.
Phat et al., (2020) [60]	LSTM	90%;	The findings revealed that the proposed approach shows promising results for real-world applications in Vietnamese datasets.

Identification through Deep Learning

Several authors have contributed to the advancement of deep learning applications to improve classification accuracy and efficiency.**Ortiz-Perez et al. (2023) [65]**and **Wadud et al. (2023) [66]** demonstrated the effectiveness of deep learning models, with Ortiz-Perez focusing on dementia detection using text modality and Wadud introducing the Deep- (Bidirectional Encoder Representations from Transformers) BERT model for offensive text classification.

Kapočiūtė-Dzikienė et al. (2022) [67],Shen et al. (2022) [68], and **Shanmugavadivel et al. (2022)[69]**each explored deep learning methods for various text classification tasks. Kapočiūtė-Dzikienė proposed a combination of fastText and CNN for news classification across multiple languages, while Shen showcased the superiority of the BERT model in status classification, and Shanmugavadivel suggested the effectiveness of the adapter-BERT model in sentiment analysis and offensive language identification.

Mehmood et al. (2022)[70]and **Chakravarthi et al. (2022) [71]** introduced novel deep learning architectures that outperformed traditional models. Mehmood explored stacked models, showing improvements over both machine learning and deep learning methods, while Chakravarthi presented a custom deep network architecture leveraging T5-Sentence embeddings.

Key findings from the survey of material available in the open literature

Table 4 shows the comparative study between several techniques applied for identification.

Table 3: Literature Review

Author	Techniques	Accuracy	Outcomes
Ortiz-Perez et al., (2023) [65]	CNN	90.36%	The text modality demonstrated superior performance in detecting dementia.
Wadud et al., (2023) [66]	Deep CNN	91.83%	The suggested Deep-BERT model outperformed all current algorithms for offensive text classification in terms of performance.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

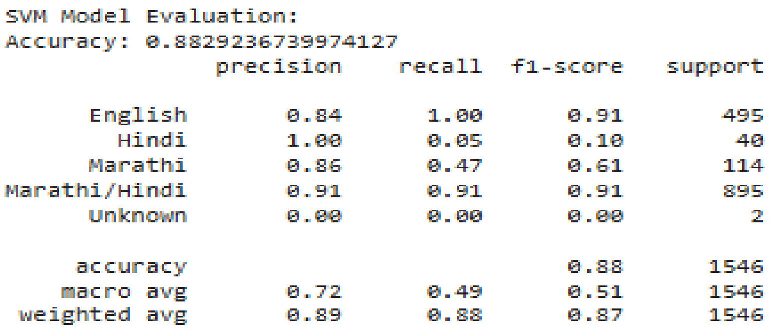
Kapočiūtė-Dzikienė et al., (2022) [67]	CNN	78.8%	The findings revealed that the combination of fastText and a CNN yielded the best performance, showcasing promising results in the classification of fake, satirical, and legitimate news across multiple languages.
Shen et al., (2022) [68]	BERT models	93%	The findings revealed that the UMLS BERT model demonstrated superior performance in classifying status.
Shanmugavadivel et al., (2022) [69]	BERT, RoBERT	79%	The findings indicate that when compared to other trained models, the adapter-BERT model shows better accuracy for sentiment analysis and objectionable language identification.
Mehmood et al., (2022) [70]	Bernoulli Naive Bayes (BNB), Logistic Regression	74.01%	According to experimental findings, the stacked model performs better than deep learning and machine learning models. with notable gains in F1 score, recall, accuracy, and precision.
Chakravarthi et al., (2022) [71]	SVM,LR, KNN	92%	The results of the study indicated that the proposed custom deep network architecture, which utilizes a concatenation of embeddings from T5-Sentence, outperformed other machine learning models.

RESULTS AND DISCUSSION

The results of the machine learning models for multilingual identification and recognition show varying levels of performance. The SVM model stands out with the highest accuracy of 88.29%, demonstrating its strong suitability for recognizing English and mixed Marathi/Hindi texts. Models like Logistic Regression, Random Forest, and Static Regression show moderate accuracy (around 62-63%), with struggles in recognizing Hindi and English samples. The XGBoost model, with the lowest accuracy of 59.96%, faces significant challenges across all languages. Overall, while SVM outperforms the others, improvements are needed for all models, especially for minority language recognition like Hindi and Marathi.

4.1 Comparison

1. SVM



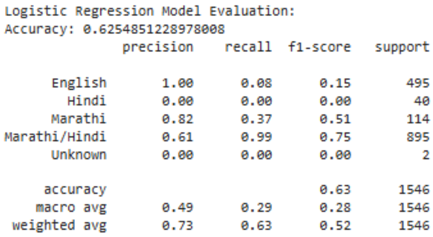
```
SVM Model Evaluation:
Accuracy: 0.8829236739974127
               precision    recall  f1-score   support

      English       0.84      1.00      0.91       495
        Hindi       1.00      0.05      0.10        40
      Marathi       0.86      0.47      0.61       114
Marathi/Hindi       0.91      0.91      0.91       895
      Unknown       0.00      0.00      0.00         2

     accuracy                           0.88      1546
    macro avg       0.72      0.49      0.51      1546
 weighted avg       0.89      0.88      0.87      1546
```

Fig 2. SVM Model Accuracy

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The SVM model performs well for both English (F1-score: 0.91) and mixed Marathi/Hindi texts (F1-score: 0.91), with an accuracy of 88.29%. It has trouble recognizing Hindi (poor recall: 0.05) and Marathi (F1-score: 0.61), even if its accuracy is respectable. Unknown instances cannot be classified by the model. The macro recall of 0.49 indicates difficulties with minority classes, even if the weighted F1-score of 0.87 indicates generally strong performance. All things considered, it works well for English and mixed texts but is less dependable for Hindi and Marathi.

2. Logistics Regression

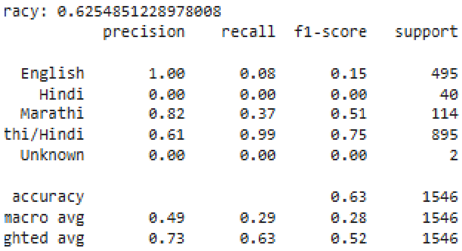


```
Logistic Regression Model Evaluation:
Accuracy: 0.6254851228978008
               precision    recall  f1-score   support

      English       1.00      0.08      0.15       495
        Hindi       0.00      0.00      0.00        40
      Marathi       0.82      0.37      0.51       114
Marathi/Hindi       0.61      0.99      0.75       895
      Unknown       0.00      0.00      0.00         2

     accuracy                           0.63      1546
    macro avg       0.49      0.29      0.28      1546
 weighted avg       0.73      0.63      0.52      1546
```

Fig 3. Logestic Regression Accuracy

With a 62.55% accuracy rate, 63% of samples are properly classified by the Logistic Regression model. It fails to completely categorize Hindi (F1-score: 0.00) and performs badly for English, with low recall (0.08) despite excellent accuracy. With an F1-score of 0.51 for Marathi, it exhibits poor recall (0.37) and intermediate accuracy (0.82). It has a higher recall (0.99) but a poorer accuracy (0.61) for mixed Marathi/Hindi (F1-score: 0.75). Cases that are unknown are not categorized. With the exception of mixed Marathi/Hindi, it has trouble with most languages overall.

3. Static Regression

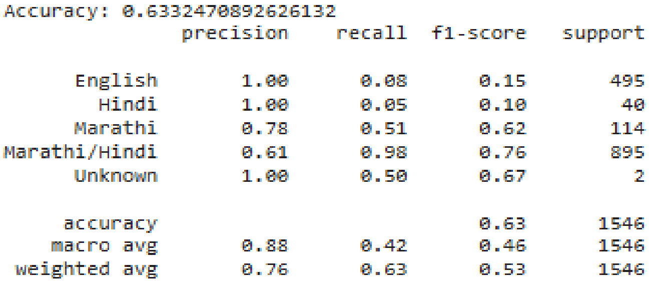


```
racy: 0.6254851228978008
          precision    recall  f1-score   support

  English       1.00      0.08      0.15       495
    Hindi       0.00      0.00      0.00        40
  Marathi       0.82      0.37      0.51       114
thi/Hindi       0.61      0.99      0.75       895
  Unknown       0.00      0.00      0.00         2

 accuracy                           0.63      1546
macro avg       0.49      0.29      0.28      1546
ghted avg       0.73      0.63      0.52      1546
```

Fig. 4. Static Regression Accuracy

With a 62.55% accuracy rate, 63% of the samples were properly classified by the Logistic Regression model. It has trouble classifying Hindi (F1-score: 0.00) and poor recall (0.08) in English, even with flawless accuracy. Marathi's F1-score is 0.51, with a moderate accuracy (0.82) and poor recall (0.37). Given its high recall (0.99) and poor accuracy (0.61), the model's F1-score of 0.75 indicates that it works well for mixed Marathi/Hindi. Unknown instances are not categorized at all. The weighted F1-score (0.52) and low macro recall (0.29) indicate its difficulties with the majority of courses.

4. Random Forest



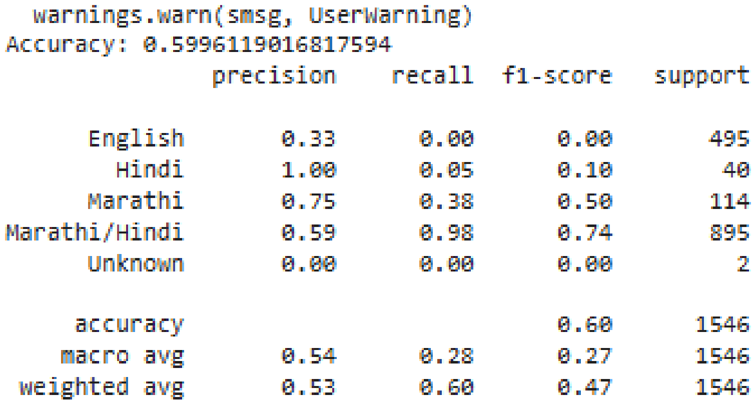
```
Accuracy: 0.6332470892626132
               precision    recall  f1-score   support

      English       1.00      0.08      0.15       495
        Hindi       1.00      0.05      0.10        40
      Marathi       0.78      0.51      0.62       114
Marathi/Hindi       0.61      0.98      0.76       895
      Unknown       1.00      0.50      0.67         2

     accuracy                           0.63      1546
    macro avg       0.88      0.42      0.46      1546
 weighted avg       0.76      0.63      0.53      1546
```

Fig 5. Random forest Accuracy

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With a weighted F1-score of 0.53 and an accuracy of 63.32%, the Random Forest model exhibits variable performance across classes. Low recall (0.08) results in a poor F1-score of 0.15, showing difficulties in detecting English samples, even while English predictions are correct (precision: 1.00). Hindi, on the other hand, has an F1-score of 0.10 because to its faultless accuracy and very poor recall (0.05). Marathi is the best-performing class, with a slightly greater recall (0.51) and an F1-score of 0.62.

2. XGBoost



```
  warnings.warn(smsg, UserWarning)
Accuracy: 0.5996119016817594
               precision    recall  f1-score   support

      English       0.33      0.00      0.00       495
        Hindi       1.00      0.05      0.10        40
      Marathi       0.75      0.38      0.50       114
Marathi/Hindi       0.59      0.98      0.74       895
      Unknown       0.00      0.00      0.00         2

     accuracy                           0.60      1546
    macro avg       0.54      0.28      0.27      1546
 weighted avg       0.53      0.60      0.47      1546
```

Fig. 6 XGBoost Accuracy

The XGBoost model achieves 59.96% accuracy but struggles across classes. It performs poorly for English (F1-score: 0.00) due to very low recall (0.00). For Hindi, despite perfect precision (1.00), low recall (0.05) results in an F1-score of 0.10. Marathi shows moderate performance with an F1-score of 0.50 (recall: 0.38). The model performs best for Marathi/Hindi mixed texts, achieving high recall (98%) and an F1-score of 0.74, indicating better recognition for combined classes.

3. Comparison Model Performance

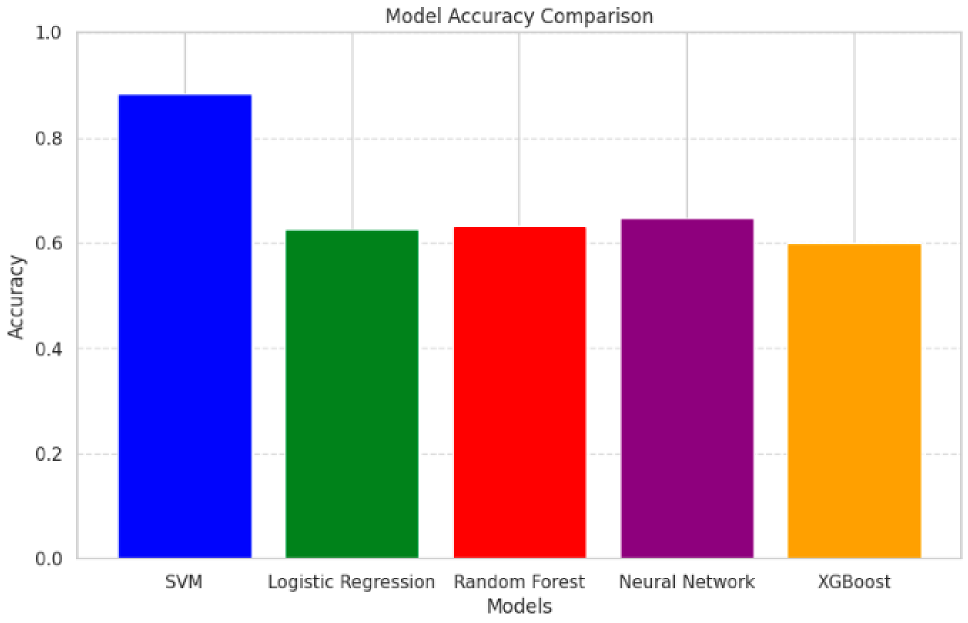


Fig 7. Comparison Models Accuracy

The SVM model achieves the highest accuracy at 88.29%, outperforming all other models. It performs particularly well for English and mixed Marathi/Hindi texts. The Random Forest model follows with an accuracy of 63.32%, showing moderate performance but struggles with languages like English and

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Hindi. Both the Logistic Regression and Static Regression models have an accuracy of 62.55%, with similar moderate performance and difficulties in recognizing Hindi and English. The XGBoost model has the lowest accuracy at 59.96%, performing poorly across most classes, especially English and Hindi. Overall, SVM provides the best accuracy by far.

Table 5. Comparison Table

Model	Accuracy (%)	Performance Summary
SVM	88.29	SVM excels in English recognition but struggles with Hindi and Marathi due to imbalance.
Logistic Regression	62.55	Logistic Regression shows bias towards dominant classes, with poor performance for English.
Static Regression	62.55	Similar to Logistic Regression, it struggles with Hindi and English recognition.
Random Forest	63.32	Random Forest performs moderately but struggles with English and Hindi, though better at Marathi.
XGBoost	59.96	XGBoost has poor performance in English and Hindi, but performs reasonably for Marathi.

The work demonstrates that traditional multilingual text recognition methods can handle simple cases with moderate success. However, challenges arise when dealing with complex images, particularly when text is embedded in noisy or low-contrast backgrounds. The performance of these methods is limited by the complexity of the text and image quality.

Strengths: The ability of existing methods to process and recognize clear and standardized text in multilingual settings.

Weaknesses: Difficulty in handling text complexity, varying image quality, and accurate language identification in complex scenes.

CONCLUSION

With a weighted F1-score of 0.87 and an overall accuracy of 88.29%, SVM fared better than the other machine learning models evaluated for multilingual identification and recognition. It demonstrated a notable performance disparity as a result of the class imbalance, performing well when recognizing English samples but poorly when identifying Hindi data. The unknown class completely failed, whereas Marathi did somewhat, highlighting the need for better feature engineering and dataset balancing. Despite their moderate accuracy (62.55% and 63.32%, respectively), Random Forest and Logistic Regression had trouble maintaining class balance, producing subpar results for Hindi and the unknown class. Bias was apparent even though the Marathi/Hindi mixed class performed better. Since SVM's shortcomings indicate the need for more improvement, the research emphasizes the need of resolving class imbalance, optimizing features, and investigating cutting-edge methodologies to enhance performance, particularly for underrepresented classes.

FUTURE SCOPE

Overcoming the noted constraints and improving the efficacy of the multilingual identification system are the main goals of this study's future scope. Accuracy will increase if class imbalance is addressed by dataset rebalancing and improved feature engineering, particularly for underrepresented classes like Hindi and the unknown class. While hyper parameter modification for models like Neural Networks and XGBoost may further improve their accuracy, ensemble learning approaches can be used to produce more stable and balanced performance across all classes. The technology will be more applicable in multilingual areas like India if it is expanded to accommodate more languages. Furthermore, creating a framework for real-time language recognition would make it feasible to use it practically to activities like querying, indexing, and text categorization. The system's capacity to manage intricate patterns in multilingual text data may also be improved by investigating deep learning techniques like transformers and LSTM networks, opening the door to a more precise, well-rounded and scalable solution.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

[1] AlShuweihi, Mohamed, Said A. Salloum, and Khaled Shaalan. "Biomedical corpora and natural language processing on clinical text in languages other than English: a systematic review." Recent advances in intelligent systems and smart applications (2021): 491-509.

[2] Mujahid, Muhammad, Ernesto Lee, Furqan Rustam, Patrick Bernard Washington, Saleem Ullah, Aijaz Ahmad Reshi, and Imran Ashraf. "Sentiment analysis and topic modeling on tweets about online education during COVID-19." Applied Sciences 11, no. 18 (2021): 8438.

[3] Ahmad, Gazi Imtiyaz, Jimmy Singla, Ali Anis, Aijaz Ahmad Reshi, and Anas A. Salameh. "Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus: A comprehensive review." International Journal of Advanced Computer Science and Applications 13, no. 2 (2022).

[4] Jauhiainen, Tommi, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. "Automatic language identification in texts: A survey." Journal of Artificial Intelligence Research 65 (2019): 675-782.

[5] Omayio, Enock Osoro, Indu Sreedevi, and Jeebananda Panda. "Language-Based Text Categorization: A Survey." Digital Techniques for Heritage Presentation and Preservation (2021): 11-36.

[6] Chen, Zhuo, Fei Yin, Xu-Yao Zhang, Qing Yang, and Cheng-Lin Liu. "MuLTReNets: Multilingual text recognition networks for simultaneous script identification and handwriting recognition." Pattern Recognition 108 (2020): 107555.

[7] Babhulgaonkar, Arun, and Shefali Sonavane. "Language identification for multilingual machine translation." In 2020 International Conference on Communication and Signal Processing (ICCSP), pp. 401-405. IEEE, 2020.

[8] Yuan, Tai-Ling, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. "A large chinese text dataset in the wild." Journal of Computer Science and Technology 34 (2019): 509-521.

[9] Wang, Yuxin, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. "From two to one: A new scene text recognizer with visual language modeling network." In Proceedings of the IEEECVF International Conference on Computer Vision, pp. 14194-14203. 2021.

[10] Wang, Tianwei, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. "Decoupled attention network for text recognition." In Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, pp. 12216-12224. 2020.

[11] Xiao, Zheng, Zhenyu Nie, Chao Song, and Anthony Theodore Chronopoulos. "An extended attention mechanism for scene text recognition." Expert Systems with Applications 203 (2022): 117377.

[12] Yu, Deli, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. "Towards accurate scene text recognition with semantic reasoning networks." In Proceedings of the IEEECVF conference on computer vision and pattern recognition, pp. 12113-12122. 2020.

[13] Yue, Xiaoyu, Zhanghui Kuang, Chenhao Lin, Hongbin Sun, and Wayne Zhang. "Robustscanner: Dynamically enhancing positional clues for robust text recognition." In European Conference on Computer Vision, pp. 135-151. Cham: Springer International Publishing, 2020.

[14] Yan, Ruijie, Liangrui Peng, Shanyu Xiao, and Gang Yao. "Primitive representation learning for scene text recognition." In Proceedings of the IEEECVF conference on computer vision and pattern recognition, pp. 284-293. 2021.

[15] Zhang, Xinyun, Binwu Zhu, Xufeng Yao, Qi Sun, Ruiyu Li, and Bei Yu. "Context-based contrastive learning for scene text recognition." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 3, pp. 3353-3361. 2022.

[16] Ke, Wenjun, Qingzhi Hou, Yutian Liu, Xinyue Song, and Jianguo Wei. "SARN: Script-Aware Recognition Network for scene multilingual text recognition." Expert Systems with Applications 250 (2024): 123753.

[17] Mehrotra, Kapil, Manish Kumar Gupta, and Karan Khajuria. "Collaborative deep neural network for printed text recognition of indian languages." In 2019 Fifth International Conference on Image Information Processing (ICIIP), pp. 252-256. IEEE, 2019.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[18] Najahan Binti Mohd Rashidi, Atina, Pantea Keikhosrokiani, Moussa Pourya Asl, and Henry Oinas-Kukkonen. "Computational analysis of dystopian elements in the partition fiction: A machine learning approach to the indian English novels." (2024).

[19] Gholami, Fatemeh, Zahed Rahmati, Alireza Mofidi, and Mostafa Abbaszadeh. "On Enhancement of Text Classification and Analysis of Text Emotions Using Graph Machine Learning and Ensemble Learning Methods on Non-English Datasets." Algorithms 16, no. 10 (2023): 470.

[20] Salh, Dana Abubakr, and Rebwar Mala Nabi. "Kurdish Fake News Detection Based on Machine Learning Approaches." Passer journal of basic and applied sciences 5, no. 2 (2023): 262-271.

[21] Alfartosy, Hadeel H., and Hussein K. Khafaji. "A New Feature Extraction, Reduction, and Classification Method for Documents Based on Fourier Transformation." International Journal of Intelligent Engineering & Systems 16, no. 5 (2023).

[22] Das, Rajesh Kumar, Mirajul Islam, Md Mahmudul Hasan, Sultana Razia, Mocksidul Hassan, and Sharun Akter Khushbu. "Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models." Heliyon 9, no. 9 (2023).

[23] Al-onazi, Badriyya B., Najm Alotaibi, Jaber S. Alzahrani, Hussain Alshahrani, Mohamed Ahmed Elfaki, Radwa Marzouk, Mahmoud Othman, and Abdelwahed Motwakel. "Modified Dragonfly Optimization with Machine Learning Based Arabic Text Recognition." Computers, Materials & Continua 76, no. 2 (2023).

[24] Wasim, Muhammad, Sehrish Munawar Cheema, and Ivan Miguel Pires. "Normalized effect size (NES): a novel feature selection model for Urdu fake news classification." PeerJ Computer Science 9 (2023): e1612.

[25] Fkih, Fethi, Mohammed Alsuhaibani, Delel Rhouma, and Ali Mustafa Qamar. "Novel Machine Learning–Based Approach for Arabic Text Classification Using Stylistic and Semantic Features." CMC-COMPUTERS MATERIALS & CONTINUA 75, no. 3 (2023): 5871-5886.

[26] Dai, Qian. "Construction of English and American literature corpus based on machine learning algorithm." Computational Intelligence and Neuroscience 2022 (2022).

[27] Tjandra, Andros, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh, Alexis Conneau, Alexei Baevski, Assaf Sela, Yatharth Saraf, and Michael Auli. "Improved language identification through cross-lingual self-supervised learning." In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6877-6881. IEEE, 2022.

[28] Shashirekha, Hosahalli Lakshmaiah, Fazlourrahman Balouchzahi, Mudoor Devadas Anusha, and Grigori Sidorov. "CoLI-machine learning approaches for code-mixed language identification at the word level in Kannada-English texts." arXiv preprint arXiv:2211.09847 (2022).

[29] Polat, Huseyin, and Mesut Korpe. "Estimation of demographic traits of the deputies through parliamentary debates using machine learning." Electronics 11, no. 15 (2022): 2374.

[30] Barua, Adrita, Omar Sharif, and Mohammed Moshiul Hoque. "Multi-class sports news categorization using machine learning techniques: resource creation and evaluation." Procedia Computer Science 193 (2021): 112-121.

[31] To, Huy Quoc, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. "Gender prediction based on vietnamese names with machine learning techniques." In Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval, pp. 55-60. 2020.

[32] Phat, Huu Nguyen, and Nguyen Thi Minh Anh. "Vietnamese text classification algorithm using long short term memory and Word2Vec." Информатика и автоматизация 19, no. 6 (2020): 1255-1279.

[33] Sarwar, Raheem, Attapol T. Rutherford, Saeed-Ul Hassan, Thanawin Rakthanmanon, and Sarana Nutanong. "Native language identification of fluent and advanced non-native writers." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 19, no. 4 (2020): 1-19.

[34] Abonizio, Hugo Queiroz, Janaina Ignacio De Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. "Language-independent fake news detection: English, Portuguese, and Spanish mutual features." Future Internet 12, no. 5 (2020): 87.

[35] Naqvi, Rizwan Ali, Muhammad Adnan Khan, Nauman Malik, Shazia Saqib, Tahir Alyas, and Dildar Hussain. "Roman Urdu news headline classification empowered with machine learning." Comput. mater. contin 65 (2020): 1221-1236.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[36] Alharbi, Adel R., and Amer Aljaedi. "Predicting rogue content and Arabic spammers on twitter." Future Internet 11, no. 11 (2019): 229.
[37] Ortiz-Perez, David, Pablo Ruiz-Ponce, David Tomás, Jose Garcia-Rodriguez, M. Flores Vizcaya-Moreno, and Marco Leo. "A Deep Learning-Based Multimodal Architecture to predict Signs of Dementia." Neurocomputing 548 (2023): 126413.
[38] Wadud, Md Anwar Hussen, Muhammad F. Mridha, Jungpil Shin, Kamruddin Nur, and Aloke Kumar Saha. "Deep-BERT: Transfer Learning for Classifying Multilingual Offensive Texts on Social Media." Computer Systems Science & Engineering 44, no. 2 (2023).
[39] Kapočiūtė-Dzikienė, Jurgita, and Askars Salimbajevs. "Comparison of Deep Learning Approaches for Lithuanian Sentiment Analysis." Baltic Journal of Modern Computing 10, no. 3 (2022): 283-294.
[40] Shen, Zitao, Dalton Schutte, Yoonkwon Yi, Anusha Bompelli, Fang Yu, Yanshan Wang, and Rui Zhang. "Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision." BMC medical informatics and decision making 22, no. Suppl 1 (2022): 88.
[41] Shanmugavadivel, Kogilavani, V. E. Sathishkumar, Sandhiya Raja, T. Bheema Lingaiah, S. Neelakandan, and Malliga Subramanian. "Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data." Scientific Reports 12, no. 1 (2022): 21557.
[42] Mehmood, Aneela, Muhammad Shoaib Farooq, Ansar Naseem, Furqan Rustam, Mónica Gracia Villar, Carmen Lili Rodríguez, and Imran Ashraf. "Threatening URDU language detection from tweets using machine learning." Applied Sciences 12, no. 20 (2022): 10342.
[43] Chakravarthi, Bharathi Raja. "Hope speech detection in YouTube comments." Social Network Analysis and Mining 12, no. 1 (2022): 75.
[44] Das, Amit Kumar, Abdullah Al Asif, Anik Paul, and Md Nur Hossain. "Bangla hate speech detection on social media using attention-based recurrent neural network." Journal of Intelligent Systems 30, no. 1 (2021): 578-591.
[45] Singh, Gundeep, Sahil Sharma, Vijay Kumar, Manjit Kaur, Mohammed Baz, and Mehedi Masud. "Spoken language identification using deep learning." Computational Intelligence and Neuroscience 2021 (2021).
[46] Yasir, Muhammad, Li Chen, Amna Khatoon, Muhammad Amir Malik, and Fazeel Abid. "Mixed script identification using automated DNN hyperparameter optimization." Computational intelligence and neuroscience 2021 (2021).
[47] Baba, Mitsuru, Tomoya Imamura, Naoto Hoshikawa, Hirotaka Nakayama, Tomoyoshi Ito, and Atsushi Shiraki. "Development of a multilingual digital signage system using a directional volumetric display and language identification." OSA Continuum 3, no. 11 (2020): 3187-3196.
[48] Vadavalli, Adilakshmi, and R. Subhashini. "Deep Learning based truth discovery algorithm for research the genuineness of given text corpus." International Journal of Recent Technology and Engineering (2019).
[49] Zhao, Fen, Penghua Li, Yuanyuan Li, Jie Hou, and Yinguo Li. "Semi-supervised convolutional neural network for law advice online." Applied Sciences 9, no. 17 (2019): 3617.
[50] Jamatia, Anupam, Amitava Das, and Björn Gambäck. "Deep learning-based language identification in English-Hindi-Bengali code-mixed social media corpora." Journal of Intelligent Systems 28, no. 3 (2019): 399-408.
[51] Mednini, Latifa, Zouhaira Noubigh, and Mouna Damak Turki. "Natural language processing for detecting brand hate speech." Journal of Telecommunications and the Digital Economy 12, no. 1 (2024): 486-509.
[52] Hasan, Mahmud, Labiba Islam, Ismat Jahan, Sabrina Mannan Meem, and Rashedur M. Rahman. "Natural language processing and sentiment analysis on bangla social media comments on Russia–Ukraine war using transformers." Vietnam Journal of Computer Science 10, no. 03 (2023): 329-356.
[53] Jain, Vipin, and Kanchan Lata Kashyap. "Ensemble hybrid model for Hindi COVID-19 text classification with metaheuristic optimization algorithm." Multimedia Tools and Applications 82, no. 11 (2023): 16839-16859.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[54] Qureshi, Muhammad Aasim, Muhammad Asif, Saira Anwar, Umar Shaukat, Muhammad Adnan Khan, and Amir Mosavi. "Aspect Level Songs Rating Based Upon Reviews in English." Computers, Materials & Continua 74, no. 2 (2023).

[55] Zhao, Yunsong, Bin Ren, Wenjin Yu, Haijun Zhang, Di Zhao, Junchao Lv, Zhen Xie et al. "Construction of an Assisted Model Based on Natural Language Processing for Automatic Early Diagnosis of Autoimmune Encephalitis." Neurology and Therapy 11, no. 3 (2022): 1117-1134.

[56] Adipradana, Ryan, Bagas Pradipabista Nayoga, Ryan Suryadi, and Derwin Suhartono. "Hoax analyzer for Indonesian news using RNNs with fasttext and glove embeddings." Bulletin of Electrical Engineering and Informatics 10, no. 4 (2021): 2130-2136.

[57] Fan, Xiaoming, Tuo Shi, Jiayan Cai, and Binjun Wang. "CHARM: An Improved Method for Chinese Precoding and Character-Level Embedding." IEEE Access 9 (2021): 129539-129551.

[58] Tokgoz, Meltem, Fatmanur Turhan, Necva Bolucu, and Burcu Can. "Tuning language representation models for classification of Turkish news." In 2021 International symposium on electrical, electronics and information engineering, pp. 402-407. 2021.

[59] Sarthak, Shikhar Shukla, and Govind Mittal. "Spoken language identification using convnets." In Ambient Intelligence: 15th European Conference, AmI 2019, Rome, Italy, November 13–15, 2019, Proceedings 15, pp. 252-265. Springer International Publishing, 2019.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.