VIVA-VOCE

"BINARY MULTILINGUAL MACHINE-GENERATED TEXT DETECTION"

Batch Number: CSE -32

Roll Number	Student Name
20211CSE0143	G.Jahnavi
20211CSE0144	U.Yashaswini
20211CSE0169	J.Meghana
20211CSE0133	K.Kowshik Narayan

Under the Supervision of,

Mr.JayaChandran Arumugam
Professor
School of Computer Science and Engineering
Presidency University

Name of the Program: computer Science & engineering
Name of the HoD: Dr .Asif Mohammed
Name of the Program Project Coordinator: Sandeep Albert Mathias
Name of the School Project Coordinators: Dr. Sampath A K / Dr. Abdul Khadar A / Mr. Md Ziaur Rahman

Introduction

- Binary Multilingual Machine-Generated Text Detection aims to develop a highly accurate system for distinguishing between human-written and machine-generated text across multiple languages.
- With the rise of automated content creation tools, ensuring content authenticity has become crucial in domains like security and journalism.
- This project leverages advanced machine learning and deep learning models, including Random Forest, LSTM, BERT, Decision Tree, and Logistic Regression, trained on a dataset of over 960,000 samples with diverse features like language, source, and model type. By providing a reliable, multilingual detection tool, the project enhances digital content integrity and supports global content verification efforts.

Literature Review

- Han and Kim (2020) reviewed the transition from traditional methods like TF-IDF and SVM to deep learning techniques such as CNNs and transformers, emphasizing the benefits and challenges of deep models.
- Wu et al. (2022) demonstrated the superiority of transformer models like BERT in multilingual classification, effectively capturing semantic nuances.
- Al-Hadhrami et al. (2020) examined algorithms for fake news detection, comparing traditional models with LSTM and BERT, highlighting accuracy challenges with imbalanced datasets.
- Li et al. (2022) introduced a multi-view learning framework that enhances cross-lingual classification accuracy by combining linguistic features.
- Conneau et al. (2020) presented XLM for cross-lingual language pretraining, emphasizing the importance of handling diverse syntax and cultural differences for effective multilingual NLP.

Research Gaps Identified

The review of text classification studies highlights key limitations across various approaches.

- Han and Kim (2020) emphasized high computational demands and limited interpretability in deep learning models.
- Wu et al. (2022) noted challenges with low-resource languages and reliance on pre-trained models.
- Al-Hadhrami et al. (2020) identified issues like dataset imbalance and over-reliance on deep learning for fake news detection.
- Li et al. (2022) pointed out complex multi-view integration and high resource needs in cross-lingual classification. Across multiple studies, common gaps include limited support for rare languages, scalability issues, and vulnerability to adversarial samples, indicating the need for more efficient and adaptable models for multilingual text classification.

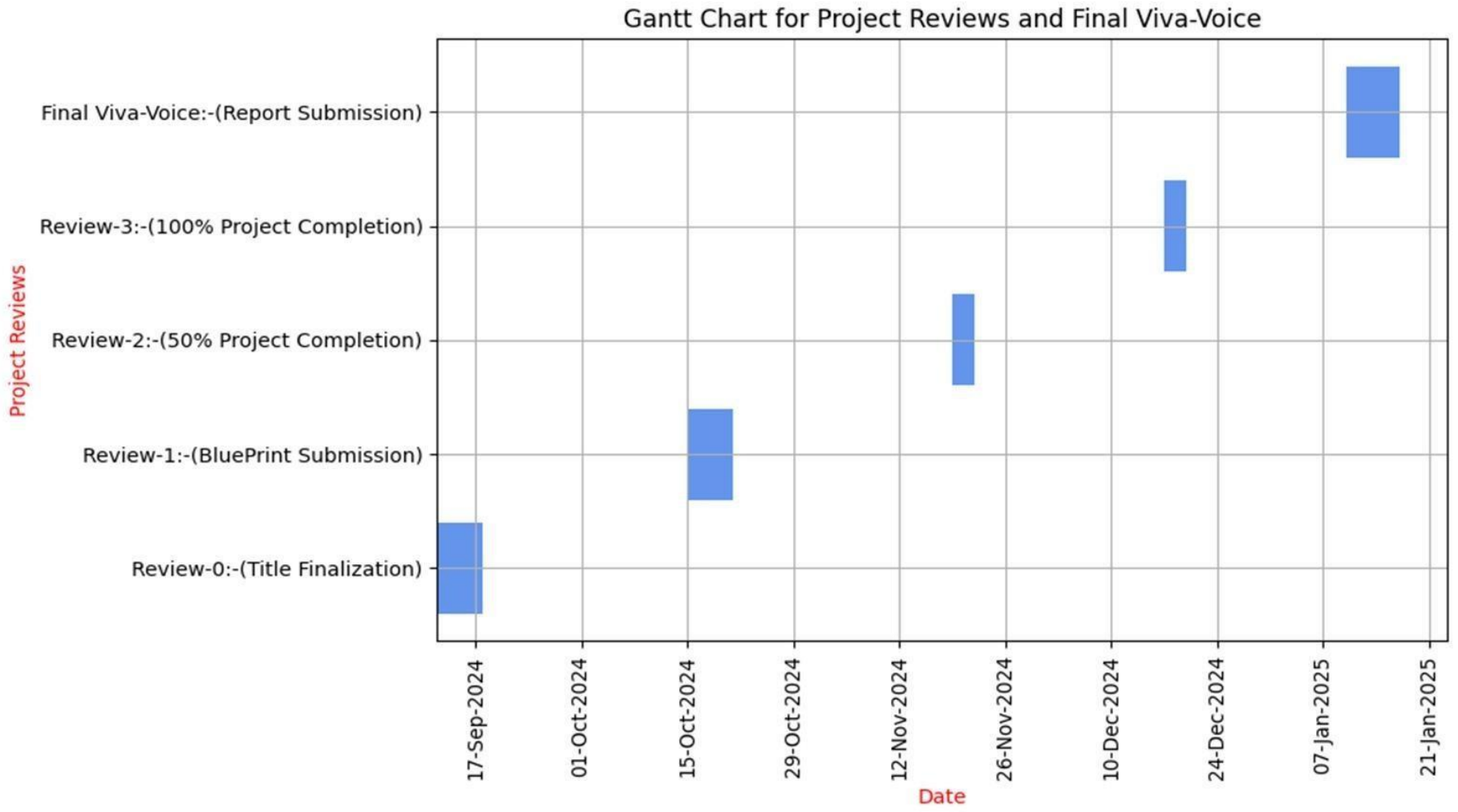
Proposed Method

- The binary multilingual text recognition system aims to detect text in images across multiple languages and scripts using a structured approach.
- The dataset preparation involves collecting diverse multilingual text samples, data augmentation, and balanced dataset splitting for training, validation, and testing.
- The model design uses lightweight CNN architectures like EfficientNet and YOLO-Tiny for efficient feature extraction, followed by binary classification with a sigmoid activation layer.
- Training strategies incorporate binary cross-entropy loss, Adam optimizer, and regularization techniques to prevent overfitting. Performance is evaluated using accuracy, precision, recall, and ROC-AUC metrics.
- Deployment optimization includes model compression for edge devices, ensuring real-time performance with frameworks like TensorFlow Lite and ONNX.

Objectives

- The binary multilingual text detection system aims to identify text presence in images across multiple languages and scripts, serving as a foundation for tasks like OCR, language translation, and content moderation.
- It focuses on high accuracy by minimizing false positives and negatives while being language-agnostic to support diverse scripts without requiring separate models.
- The system ensures robustness against variations in image quality, lighting, and complex backgrounds, while also handling rotated and distorted text.
- Real-time processing is prioritized for applications like AR and live video, balancing speed and accuracy. Designed for scalability, it supports the easy integration of new languages and evolving text styles.

Timeline of Project



Outcomes / Results Obtained

- This binary multilingual text detection system aims to provide high accuracy in identifying text across a wide range of languages and scripts.
- By focusing on binary classification, it ensures reliable text detection with minimal errors, even in challenging conditions like noisy, blurred, or occluded text.
- The system is designed for efficiency, requiring low computational resources, making it suitable for real-time applications on resource-constrained devices.
- It offers broad applicability across industries, from content moderation and document automation to augmented reality and video analysis, while also promoting linguistic inclusivity by supporting diverse languages and scripts.

Conclusion

- In conclusion, this project addresses the growing challenge of distinguishing between human-authored and machine-generated content in a multilingual setting. By combining traditional machine learning techniques such as Logistic Regression, Decision Tree, and Random Forest with advanced deep learning models like LSTM and BERT, the system leverages the strengths of both approaches to enhance detection accuracy and efficiency.
- With a diverse dataset covering languages such as English, Indonesian, German, and Russian, the system is designed to provide scalable and reliable machine-generated text detection. Ultimately, this project contributes to the development of content verification tools, fostering trust in digital information and promoting a more secure and credible digital environment across different linguistic and cultural contexts

References

- Han, T. R., & Kim, J. H. (2020). "A Survey on Text Classification: From Shallow to Deep Learning." IEEE Access, 8, 24430-24448.
- Wu, Y., Wang, X., & Xu, X. (2022). "Multilingual Text Classification with Transformer Models." IEEE Transactions on Knowledge and Data Engineering, 34(6), 1019-1032.
- Al-Hadhrami, F., Idris, M. I., & Al-Kahtani, A. (2020). "A Review of Text Classification Algorithms for Detecting Fake News." IEEE Access, 8, 134055-134070.
- Li, J., Liu, Z., & Wang, H. (2022). "Cross-lingual Text Classification Using Multi-View Learning." IEEE Transactions on Neural Networks and Learning Systems, 33(4), 1607-1620.
- Sharma, N., Singh, S. S., & Kumar, V. P. (2022). "Deep Learning for Multilingual Text Classification: A Comparative Study." IEEE Transactions on Emerging Topics in Computing, 10(2), 160-172.

Publication Details

Site: Journal of Xidian University

Thank You