

## Semantic Metadata-enhanced Deep Learning Techniques for Duplicate Bug Report Detection

Sandunika D.M.N.<sup>1\*</sup>, Herath G.A.C.A.<sup>2</sup>

<sup>1</sup>Department of Software Engineering, Faculty of Computing,  
Sabaragamuwa University of Sri Lanka, Sri Lanka

<sup>2</sup>Department of Computing and Information Systems, Faculty of Computing,  
Sabaragamuwa University of Sri Lanka, Sri Lanka

\*dmnsandunika@std.appsc.sab.ac.lk

Duplication of bug reports poses a significant impact on software development efficiency with 12-25% of bugs report being duplicated on large projects. Manual identification techniques are both time consuming and prone to error. This study examines whether semantic metadata that encodes content-level similarities together with simplified deep learning architectures can be more effective than relying on complex models. We aim to examine existing practices and determine weaknesses, research and prove semantic content-based metadata to be useful in robust identification, structure and test various deep learning models to define the best methods to use to achieve higher levels of performance. Existing studies reveal that machine learning methods outperform traditional information-retrieval techniques by a significant margin, while deep learning approaches achieve higher accuracy but often suffer from limited feature diversity. We collected 535,477 of quality pairs with a 70%-10%-20% split for training, validation, and testing. Our feature engineering on DistilBERT yielded a mean correlation of 0.1888, surpassing traditional metadata. We compared 6 architectures, namely, LSTM, CNN, LSTM+Metadata, Hybrid+Attention, Hybrid+Attention with Metadata, and proposed LSTM+CNN+Metadata, evaluated through accuracy, precision, recall, F1-score, and AUC-ROC metrics. The proposed architecture resulted 92.53% F1-score, which is better than complex attention-based models. Contextual processing is proven to be better, as LSTM performs higher than CNN. This paper highlights that the quality of features is more critical for model performance than model complexity, presenting a cost-effective, accurate, and scalable method for automated duplicate bug detection in real-world applications.

**Keywords:** *Deep Learning, Duplicate Bug Detection, LSTM Networks, Semantic Metadata, Software Maintenance*