***Abstract No: 168*** ***ICT, Mathematics, and Statistics***

# CLASSIFICATION OF SINHALA NEWS USING MACHINE LEARNING APPROACHES

**S.P.D.H. Nawarathna* and S. Mahesan**

*Department of Computer Science, Faculty of Science, University of Jaffna, Jaffna, Sri Lanka.*
**nawarathnadeshan@gmail.com*

With the advent of internet technology, the popularity of Sinhala text-based news portals has witnessed a significant escalation. To aid users in efficiently locating news articles relevant to their interests, this study introduces a systematic approach for classifying Sinhala news headlines, leveraging machine learning methodologies. The system curates a dataset of 25,400 news articles from Sinhala news websites, meticulously labelled for training and evaluation purposes. It explores various text embedding techniques, including term frequency-inverse document frequency, Word2Vec, and FastText, while employing classification algorithms such as support vector machines, Naive Bayes, Logistic Regression with multi-class classification, and long short-term memory (LSTM) networks. The experimental outcomes underscore that the most effective combination for classifying Sinhala news headlines is the integration of FastText and LSTM, achieving an impressive accuracy rate of 93.8% for news headlines alone and 95.8% when applied to a mixed dataset encompassing both news headlines and news content. Furthermore, the LSTM classifier demonstrates its ability to capture long-term dependencies within the text, a crucial factor in ensuring the precise classification of Sinhala news headlines. This research highlights that the LSTM + FastText combination yields superior accuracy in classifying Sinhala news, thus making it a noteworthy approach for this purpose.

**Keywords**: FastText, Logistic regression**,** Text embedding