5th International Conference on Advanced Research in Computing (ICARC) 2025
Graduate Colloquium

Paper ID:72

A Comparative Analysis of Deep Learning Algorithms for Formality Classification in Texts Using Linguistic Features

Karunarathna K.M.G.S.[1*], Rupasingha R.A.H.M.[2], Kumara B.T.G.S.[3]

[1] Faculty of Graduate Studies, [2]Faculty of Social Sciences and Languages, [3]Faculty of Computing, [1,2,3]Sabaragamuwa University of Sri Lanka

[*]gayathrisarangika599@gmail.com

Because of the wide variety of formal and informal writing styles brought about by the rapid growth of digital communication, the classification of documents based on it becomes a challenging task. Using a variety of variables, this work seeks to increase the accuracy of formality classification algorithms. Grammar, vocabulary, punctuation, and sentence structure are some stylistic components that define various writing styles, and traditional approaches have trouble distinguishing between them. Differentiating between formal and informal language is becoming increasingly important in applications such as research papers, legal documents, informal letters, NEWS, etc. The objective of this approach is to use linguistic features, examines how well deep learning algorithms classify documents as formal and informal. The study collected dataset of 5,000 text samples. The text files contained 2500 formal letters, news items as formal documents, and remaining are personal blogs, personal letters as informal documents. Next pre-processed all data using stop word removal, lemmatization, tokenization and lowercasing. Formal and informal categories which include pronouns, grammar, vocabulary, slang, acronyms, language and initialisms seven linguistic features were targeted for this study and those features are extracted. Then these seven features are combined to generated the feature vector for each document. The generated feature vector was applied and in order to classify documents, three deep learning models Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks are trained. Here, ANN learns nonlinear patterns in data, CNN identifies text sections, and LSTM considers word position and those are selected based on the literature review. The performance of each model is compared using different test splitting methods and cross-validation techniques. According to experimental data, the LSTM model outperforms ANN and CNN in terms of precision, recall and f-measure metrics, achieving the highest classification accuracy of 89.4% with an epoch size of 100 and a batch size of 32 with lowest error rate for Mean Absolute Error and Root Mean Squared Error. The results highlight how well LSTM can detect linguistic subtleties and offer suggestions for improving formality recognition in Natural Language Processing applications, which will help with more context-sensitive text classification.

Keywords: *Classification, Deep Learning, Formal documents, Informal documents, Linguistic Features*