Identify the Characteristics for Categorizing Documents based on Different Writing Styles

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

[1]*Faculty of Graduate Studies, Sabaragamuwa University of Sri Lanka, Sri Lanka*
[2]*Department of Economics and Statistics, Faculty of Social Sciences and Languages, Sabaragamuwa University of Sri Lanka, Sri Lanka*
[3]*Department of Computing and Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka, Sri Lanka*
[*]*gayathrisarangika599@gmail.com*

As technology advances, more people are being persuaded to use the internet to acquire information. On the internet, people may find a wide range of documents, including scholarly papers, academic books, reports, research articles, etc. However, in general, web papers are not logically organized, which makes it difficult and time-consuming to get pertinent information from a website. Therefore, a particular study was accomplished to classify the documents based on formal and informal writing styles considering their characteristics. There are linguistic variations that are specific to each style that may be used to determine if a document is formal or informal. Before creating this model perceived the characteristics of the informal and formal styles. In this study, we focused on 15 characteristics, and currently, 6 characteristics are considered namely, Colloquialism, Abbreviation, Contraction, Voice, Modal Verbs, and Phrasal Verbs. Used 5000 data sets for this experiment as formal news articles, informal letters and personal blogs. Pre-processed them using four steps, such as tokenization, stop word removal, lowercasing, and lemmatization, and used four feature extractions methods: Tf-Idf, Word2Vec, Doc2Vec, and Glove. For contraction create an algorithm to find out how many contractions are included using seven rules. Modal verbs and phrasal verbs are also counted in every document and identify the passive voice and active voice separately. And also identifies the abbreviation and colloquialism by classifying the documents using two target variables. For the classification process Artificial Neural Network (ANN) and Long Short-Term Memory (LSTM) algorithms. Considering the abbreviation characteristic doc2vec showed highest accuracy in the ANN algorithm and for the colloquialism characteristic doc2vec showed the highest accuracy in the LSTM algorithm. In this approach, six features have been completed, while the remaining nine are being worked on. Finally, planning to classify documents as formal or informal, combines all of the findings.

Keywords: *Classification, Formal writing style, Informal writing style, Linguistic variation, Machine learning*