Classifying Text In Open And Distance Education Using A Deep Learning Long Short-Term Memory Neural Network

Dursun Akaslan
Harran Üniversitesi

Abstract No: 297 - Abstract Presentation Type: Oral Presentation

---

Deep Learning is considered as a branch of machine learning that teaches computers to do what comes naturally to humans and learns from experience. Deep learning with algorithms, pretrained models and apps are becoming highly popular in our century for text classification. Text classification is described as the process of automatically tagging a textual document with most relevant categories. The natural sequence of text data is used to categorise text because a piece of text consists of a sequence of words. However, a word might be dependent on the previous or next word or words.

Several techniques are used in the classification of words such as long short-term memory (LSTM) network. LSTM was first proposed in 1997 and today it is used in the field of deep learning as an artificial recurrent neural network (RNN). An LSTM neural network is known as a type oof RNN that can learn long-term dependencies between time steps of sequence data.

The purpose of this study to classify a huge amount of text data in open and distance education using a deep learning long short-term memory neural network. The text data is obtained by a semi-structured interview associated with open and distance education and coded manually. After coding the text data obtained from interviews, narrative data in the text data will be classified with several categories. This text data will be used for training, validating and testing the deep learning LSTM neural network. MATLAB was used in the study to classify the text data in six stages with the utilization of Deep Learning and Text Analytics Toolboxes namely: importing data (1), pre-processing text data (2), Converting Document to Sequences (3), Creating and Training LSTM Network (4), Testing LSTM Network (5) and Predicting using New Data (6).

The first stage is called "importing data" and used to import the text data into MATLAB as a matrix with the purpose of removing missing text. First, the text data in our study are first imported in the MATLAB Desktop as a matrix and then rows with missing data was removed in the first stage. It is also important to note here that classes with few observations are also removed from the matrixes in this stage. Moreover, the text data is partitioned into three groups for training (70%), testing (%30) and validating (%30). The second stage aims to pre-process the text by tokenizing the text data, converting uppercase in the text data to lowercase and erasing the punctuation in the text data. A word encoding is used in the third stage to convert the text data into sequences of numeric indices. While the LSTM network architecture is defined in the third stage, it was tested in the fifth stage. Finally, the new text data is classified into categories.

The findings of this study will help individuals analyse the text data automatically with the assistance of LSTM neural network after they created the LSTM architecture. This is especially important for organizations and individuals who must constantly classify massive data in their institutions.

Keywords: LSTM, open and distance education, deep learning, neural network, artifical intelligence, machine learning