Abstract

Data is being generated in high volume and velocity in our current world, data storage and data quality is becoming an important issue. Paraphrase Identification is an important problem since it can help reduce the amount of data being generated and improve overall data quality. To solve this problem, a system needs to be built that takes two texts as input and outputs a label that asserts whether the questions are semantically similar or not. Quora Question Pairs, which consists of more than 400000 labelled samples of semantically similar and dissimilar questions, is used as a dataset. It was prepared by resolving missing data fields, data duplication and data inconsistency issues, followed by tokenization and vector embedding generation using GloVe and FastText. Class imbalance was corrected using a graph-based data augmentation strategy that introduced new similar data samples. Siamese network models are a category of Neural Network architectures that fit this problem perfectly. In past research, models like BiMPM, ESIM and CAS-LSTM have used the same architecture to push the threshold of accuracy on this problem. Siamese network models in this project are designed and trained using RNN, LSTM, GRU and Transformer units as building blocks. These models are then comprehensively evaluated using accuracy, precision, recall, F1 and AUC scores. The Siamese network model built using Transformer proved effective with an accuracy of 0.83 and F1 score of 0.77, proving that the methodology followed in this project is effective at paraphrase identification.

Keywords: paraphrase identification, siamese networks, deep learning