Time complexity analysis of the presented HAN based model for the bug report prioritization

Partially the model consists of four main components: Text Preprocessing, Tokenization, oversampling training dataset, and model building and training. The model building consists of embedding layer, bidirectional layer, attention layer, and a dense layer. Listing below the time complexities of these components for each training record in the dataset:

1. Text Preprocessing: Linear operations O(N), N= length of the text.
2. Tokenization: linear O(L), L = length of the sequence.
3. Oversampling (ADASYN): Linear - depends on the numbers of samples. (84000) in thedataset
4. Train-Test Split: Linear O(N),N = number of samples.
5. Pad Sequences: Linear O(L), L = maximum sequence length.
6. Model Building and Training:

- Embedding Layer: O(L * E), where L is the sequence length and E is the embedding dimension.
- Bidirection-GRU/RNN Layers: O(L * U), where L is the sequence length and U is the number of hidden units.
- Attention Mechanism: O(L^2) for attention computation.
- Dense Layers: O(U) for dense layers.

The net complexity for iterating through E number of epochs for training can be approximated as **O(E * (L * E + L * U + L^2 + U))**

- E is constant so Quadratic complexity.
- Evaluation - linear O(N).
- In the code(max_sequence_length) is set to 3000
- The embedding dimension (embedding_dim) is set to 100. The number of hidden units in the HAN model is set to 64.