**W11 – NLP - News Articles Classification**

| **Criteria** |
| --- |
| **Execute Programming Code Assignment**  **Complete Tasks 1-3 in the News Articles Classification Notebook** |
| **Algorithm Understanding**  **In the transformer architecture in the paper "Attention Is All You Need", how does Multi-head Attention work?**  Self-attention is a way of computing the relevance (weights) of the other words in a sentence with respect to a particular word in the sentence. And to do this for all words in the sentence. Words more relevant to the word we are analyzing will have larger scores than ‘farther away’ words.  This score is calculated by performing a dot product between query and key vectors, normalizing this value, using a softmax function and finally multiplying this by a third, value, vector. These three vectors (q, k and v) are calculated by multiplying the word embeddings by three matrices (Q, K, and V) that will be trained during the training process.  The transformer architecture includes several self-attention units that that work in parallel, and this ‘layer’ is called a multi head attention. In this way we improve the performance of the attention layer by providing more attention representations (all trainable matrices are randomly initialized). The results of all heads are ‘condense’ into a single result (to be fed to the decoder) though another trainable matrix. |
| **Interview Readiness**  **What is the main idea behind Positional Encoding?**  Since ‘position matters’ (the order of the words) to understand the meaning of sentences, it is important to include some position information in the vector encoding the word (or token). The architecture of the encoder adds a vector with the positional encoding of the word to the vector with the word embeddings.  The positional encoding technique used in transformers allows to feed information about distances between different words in sentences of different lengths with the same distance metric across all sentences. |
| **Interview Readiness**  **What is Early Stopping and why do we use it?**  When we train an ML model with some training data, the model may over fit its parameters to the training data provided. This can be detected by ‘testing’ the model with validation data not seen by the model during the previous training iteration. If the model is overfitting, the validation loss will not improve and will become significantly larger than the training loss. Overfitting becomes evident as the validation loss stop improving over the training iterations (while the training loss continues to improve). The idea of early stopping is to stop training the model when the model starts to overfit as described above, instead of waiting until the training loss stops decreasing (or changes less than a certain value). The early stopped model will be able to generalize and make reasonable predictions with previously unseen data.  . |
| **Interview Readiness**  **How would explain what a transformer model is to business stakeholders (at a high level)?**  A transformer is a model architecture that performs NLP task with higher accuracy than other NLP models and with a fraction of the computational resources that they require.  The transformer can process whole sentences at a time, instead of single words, by processing all words in a sentence in parallel. The also transformer incorporates layers with multi head attention blocks that obtain information about the relevance and interaction of all words in every sentence.  A typical transformer is composed of a encoder block and a decoder block, or either encoder blocks or decoder blocks.  . |