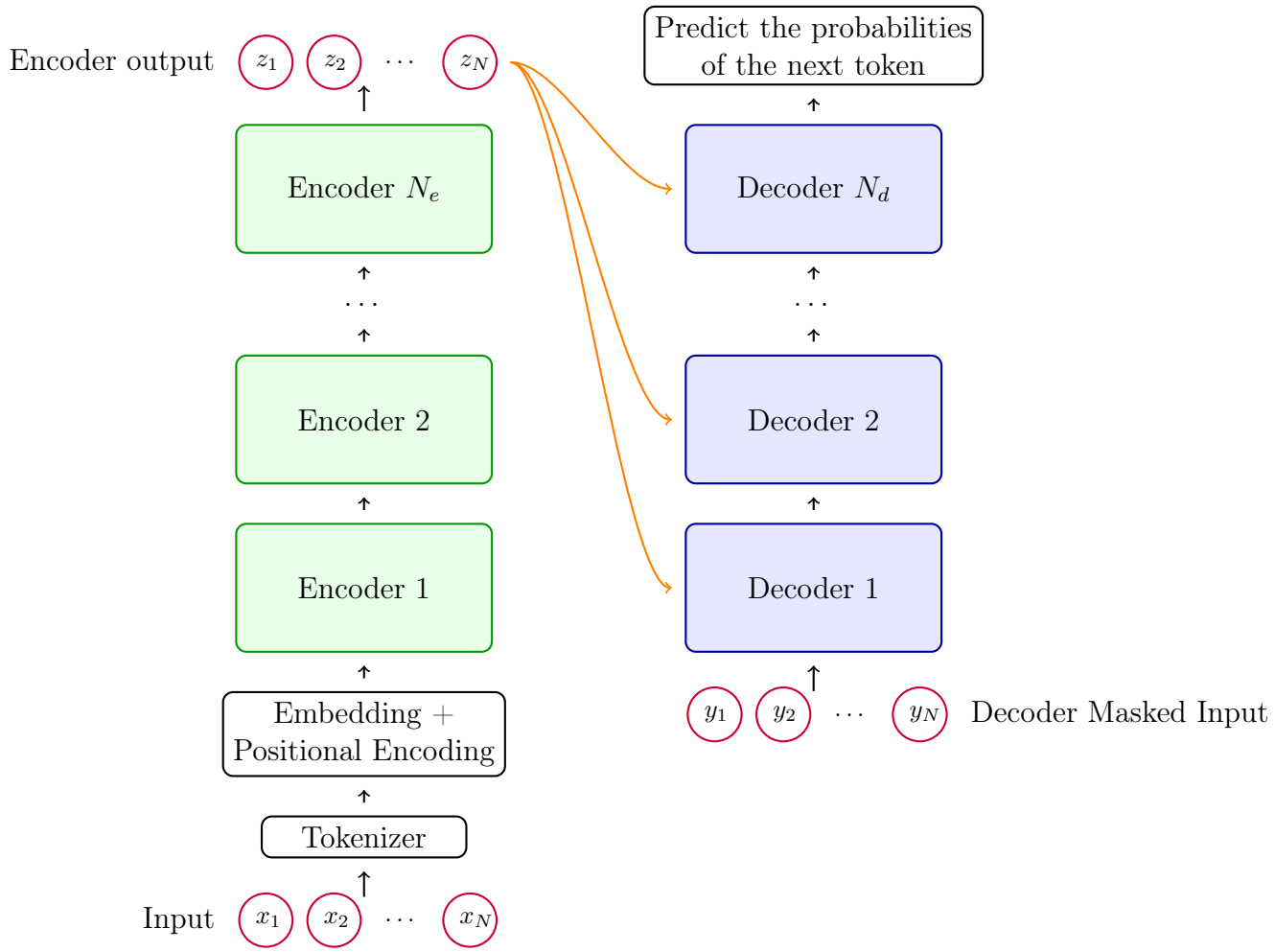
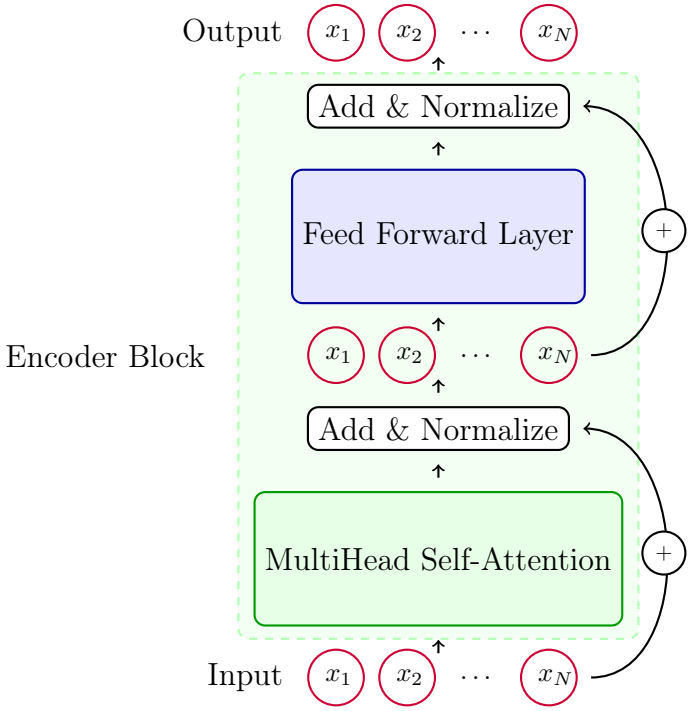
1 Structure globale de l'architecture transformer

Le schéma de l'achitecture dans le cas de la génération en mode inférence est le suivant :



2 Structure de la partie encoder

Soit N la taille de la séquence d'entrée du modèle.

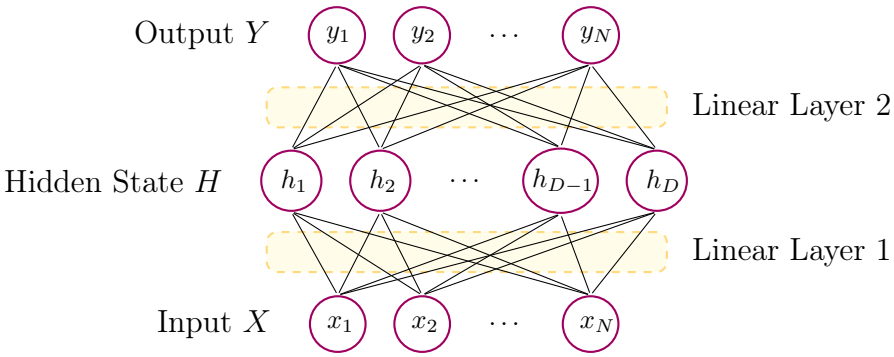


3 Schéma du Feed Forward Network :

Soit d_E la dimension des embedding des tokens. Soit N la taille de la séquence d'entrée du modèle.

Ici, X est une matrice de dimension (N, d_E) donnée en entrée du réseau, à laquelle on applique une première couche linéaire : $X \mapsto X \cdot W_1^T + B_1$ pour obtenir une matrice H de dimension (D, d_E), avec D la dimension de l'état caché du Feed Forward Network. On applique ensuite une seconde couche linéaire $H \mapsto H \cdot W_2^T + B_2$ pour obtenir la matrice de sortie Y de dimension (N, d_E).

Les matrices W_1, W_2, B_1 et B_2 sont internes au réseau et sont apprises lors de l'entraînement.



4 Structure du réseau utilisé

Je vais donc utiliser la partie encodeur du modèle BERT, qui va donc me donner une représentation vectorielle de la séquence donnée en entrée, que je vais pouvoir donner à un petit modèle (Deep) Feed Forward Classifier, dont la dernière couche linéaire aura pour sortie une dimension 1, dont je vais faire passer la sortie dans une fonction softmax afin de me donner une valeur c entre 0 et 1 que l'on pourra interpréter de la façon suivante : plus c est proche de 0, plus le texte en entrée porte un sentiment négatif, et inversement, plus c est proche de 1, plus le texte en entrée porte un sentiment positif.