**Note méthodologique : preuve de concept**

## **Dataset retenu**

*Les données retenues sont une liste de produits ainsi que leur description.*

## **Les concepts de l’algorithme récent**

*Bert est un modèle de type Transformers. Un transformer est un modèle qui fonctionne en effectuant un petit nombre constant d’étapes. À chaque étape, il applique un mécanisme d’attention pour comprendre les relations entre les mots de la phrase, quelles que soient leurs positions respectives. Prenons un exemple simple :*

*« Tu as une nouvelle souris pour ton ordinateur ? »*

*Pour déterminer le sens du mot souris, l’objet et non l’animal, le transformer va prêter attention au mot « ordinateur » et prendre une décision en une étape basée sur ça.*

*Pour permettre cela, BERT se base donc sur l’architecture des transformers, c’est-à-dire consistant en un encodeur pour lire le texte et un décodeur pour faire une prédiction. BERT se limite à un encodeur, car son objectif est de créer un modèle de représentation du langage qui sera ensuite utilisable pour des tâches de NLP. (Il permet de comprendre le langage).*

**La modélisation**

*Avant d’utiliser BERT il est crucial de faire la préparation des données suivantes :*

*Tokenisation des mots et ajout de tokens de début et de fin de phrase*

*Marqueur ajouté à chaque phrase pour les distinguer*

*Un marqueur de position est ajouté à chaque token (mots) pour indiquer sa position.*

*Après cela il faut choisir son modèle BERT. Il y en a plusieurs types de différentes tailles, il revient à l’utilisateur de choisir la complexité qui lui convient pour sa tâche. Enfin, il reste à importer le modèle et l’incorporer à votre architecture. Une fois ces étapes passées, il ne reste plus qu’à faire les prédictions !*

*BERT est un modèle de représentation du langage très puissant qui a marqué une étape importante dans le domaine du traitement automatique du langage – il a considérablement augmenté notre capacité à faire de l’apprentissage par transfert en NLP. Bert vous permettra par exemple de classifier les tweets selon le sentiment qu’ils renvoient ou encore de créer un assistant virtuel capable de répondre aux questions de façon intelligente.*

*L’indicateur de performance retenu est l’ARI.*

**Une synthèse des résultats**

*CountVectorizer : ARI = 0,4654 en 6s*

*Tf-idf : ARI = 0,4312 en 4s*

*BERT : ARI = 0,4008 en 5s*

**L’analyse de la feature importance globale et locale du nouveau modèle**

*Non applicable car il s’agit de NLP, les features sont donc les tokens extraits et ne sont pas les mêmes en fonction du texte en entrée.*

**Les limites et les améliorations possibles**

*Pour espérer améliorer les résultats, il faudra certainement envisager une meilleure préparation des données, notamment via la lemmatization / stop wording plus poussée.*