Note méthodologique :
Preuve de concept

Dataset retenu

Les données proviennent d'un site anglophone de **vente en ligne** , où des vendeurs publient des annonces avec une **photo et une description** .

Le dataset, sous format **CSV**, contient **1 050 articles** et **15 caractéristiques** . Nous nous concentrerons sur la **description des produits** et leur **catégorisation** . Il n'y a pas de données manquantes.

Nettoyage du texte

Une fonction unique et configurable effectuera plusieurs transformations :
✓ Passage en **minuscules**
✓ **Tokenisation** (suppression de la ponctuation, séparation des mots)
✓ **Filtres** optionnels :

- Stopwords (mots inutiles comme "the", "and")
- Mots rares (présents une seule fois)
- Mots de **moins de trois lettres**
- Caractères **non alphabétiques**
  ✓ **Lemmatisation** (réduction des mots à leur forme de base : "running" → "run")
  ✓ Vérification via un **dictionnaire anglais**

Ce traitement réduit le texte de **500 000 mots à 63 000** , dont **4 000 uniques** .

	description	product_category_tree	category	clean_description
814	Buy Raymond Abstract Double Blanket Blue at Rs...	["Home Furnishing >> Bed Linen >> Blankets, Qu...	Home Furnishing	buy raymond abstract doubl blanket blue onli g...
135	Prithish Working on my own Grass Ceramic Mug (...	["Kitchen & Dining >> Coffee Mugs >> Prithish ...	Kitchen & Dining	work own ceram mug price get talk with your co...
960	Cotonex Beige Cotton Kitchen Linen Set (Pack o...	["Home Furnishing >> Kitchen & Dining Linen >>...	Home Furnishing	beig cotton kitchen linen set pack price revie...
573	Gift Island AST906 Analog Watch - For Women -...	["Watches >> Wrist Watches >> Gift Island Wris...	Watches	gift island analog watch for buy gift island a...
809	Oxyglow Lacto Bleach & Fruit Massage Cream Wit...	["Beauty and Personal Care >> Combos and Kits ...	Beauty and Personal Care	bleach fruit massag cream with vitamin set pri...
362	Flipkart.com: Buy Vincent Valentine Paris Set ...	["Beauty and Personal Care >> Fragrances >> De...	Beauty and Personal Care	buy vincent valentin pari set new dark dark fi...
182	Printland PMR1902 Ceramic Mug (350 ml)\r\n ...	["Kitchen & Dining >> Coffee Mugs >> Printland...	Kitchen & Dining	ceram mug price coffe mug ador and fantast cof...
848	Buy Generix HDMI Female To Female Coupler Join...	["Computers >> Laptop Accessories >> USB Gadge...	Computers	buy femal femal coupler jointer adapt extend g...
946	Flipkart.com: Buy Organistick Silver Label Lip...	["Beauty and Personal Care >> Makeup >> Lips >...	Beauty and Personal Care	buy silver label lipstick for from price onli ...
919	Buy Anjalika Brass Laddu Gopal Showpiece - 6...	["Home Decor & Festive Needs >> Table Decor & ...	Home Decor & Festive Needs	buy brass showpiec for brass showpiec best pri...

Classification des articles

Nous utiliserons uniquement le **premier niveau** du champ **"product_category_tree"** , soit **7 catégories parfaitement équilibrées** . Cela garantit un **apprentissage optimal** du modèle.

```
category
Home Furnishing              150
Baby Care                    150
Watches                      150
Home Decor & Festive Needs   150
Kitchen & Dining             150
Beauty and Personal Care     150
Computers                    150
Name: count, dtype: int64
```

Les concepts de l'algorithme récent

Lorsqu'on veut classer des documents automatiquement, une méthode efficace consiste à utiliser un modèle basé sur un réseau de neurones. Celui-ci commence par une couche qui transforme les mots en nombres (appelée couche d'intégration), suivie d'un réseau de neurones spécial appelé **réseau neuronal convolutif** (CNN). Ensuite, on applique une couche de regroupement pour simplifier les informations, avant d'obtenir la prédiction finale.

La partie clé de ce modèle est la **couche convolutionnelle** , qui fonctionne comme une loupe glissant sur le texte pour repérer des motifs importants. Cette loupe, appelée **noyau**, peut analyser un certain nombre de mots en même temps. Par exemple, si le noyau regarde trois mots à la fois, il pourra détecter des expressions de trois mots dans le texte.

Pour améliorer encore la classification des documents, on peut utiliser **plusieurs de ces loupes en même temps** , chacune ayant une taille différente. Cela signifie qu'un noyau peut analyser des paires de mots (2 mots), un autre des groupes de trois (3 mots), et d'autres encore des séquences plus longues. Ce type de modèle s'appelle un **réseau neuronal convolutionnel multicanal** .

L'avantage de cette approche est qu'elle permet de capter des **informations à différentes échelles**. Par exemple, un petit noyau peut repérer des mots-clés, tandis qu'un plus grand peut identifier des phrases ou expressions entières. Ainsi, le modèle comprend mieux la structure du texte et prend de meilleures décisions.

Cette technique a été mise en avant la première fois en 2014 par **Yoon Kim**, un chercheur en intelligence artificielle, dans son article intitulé Convolutional Neural Networks for Sentence Classification. Il y montrait comment utiliser les réseaux neuronaux convolutifs pour analyser et classer des phrases.

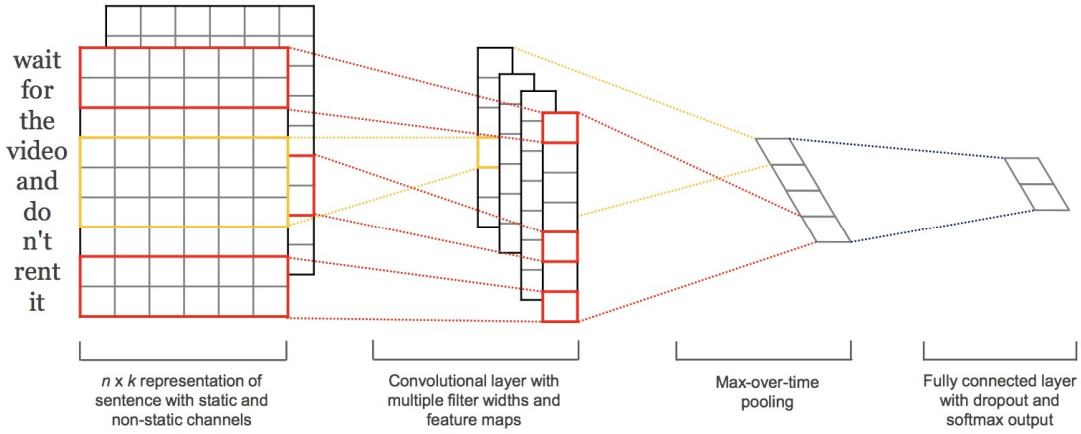
Dans son étude, Kim a testé différentes façons d'intégrer les mots dans le modèle. Il a notamment comparé une méthode où l'on fixe les représentations des mots (**statique**) et une autre où elles évoluent au fil de l'apprentissage (**dynamique**).

Toutefois, comme l'indique le didacticiel de Jason Brownlee publié sur machine learning mastery le 03 Septembre 2020, même en simplifiant l'approche et en se concentrant uniquement sur l'utilisation de **noyaux de tailles différentes** , on obtient déjà des résultats très intéressants.

Pour mieux comprendre ce fonctionnement, il est utile de **visualiser ces couches sous forme de diagramme**. Imaginez un texte sous forme d'une ligne de mots. Plusieurs fenêtres (les noyaux) glissent dessus, chacune capturant des morceaux du texte avec une taille différente. Ensuite, ces informations sont traitées et combinées pour donner la meilleure classification possible.

Ce modèle est donc un outil puissant pour analyser des textes de manière automatique. Grâce à l'utilisation de plusieurs noyaux de convolution, il capte à la fois les petits détails et les structures plus larges, permettant ainsi une **compréhension plus fine du contenu des documents** .

Cette approche est mieux comprise à l'aide d'un diagramme tiré de l'article de Kim :



La modélisation

Il faudra que notre modèle soit capable de classifier nos articles dans les sept grandes catégories de produits du site marchand.

La séparation des données

Cela signifie qu'une fois le modèle développé, nous devrons faire des prédictions sur de nouvelles descriptions textuelles. Cela nécessitera de procéder à la même préparation des données sur ces nouvelles données que sur les données d'entraînement du modèle.

Nous veillerons à ce que cette contrainte soit intégrée à l'évaluation de nos modèles en divisant les ensembles de données d'entraînement et de test avant toute préparation des données. Cela signifie que toute connaissance des données de l'ensemble de test qui pourrait nous aider à mieux préparer les données (par exemple, les mots utilisés) n'est pas disponible dans la préparation des données utilisées pour l'entraînement du modèle.

La séparation des données sera faite avec l'option de stratification pour garder des échantillons cohérents.

Cela étant dit nous utiliserons 10% des données comme ensemble de test et les 90% restant comme données d'entrainement.

La préparation

Nous tokenisons l'ensemble des données d'entrainement, ce qui nous permet de définir le vocabulaire de la couche d'embedding (couche qui transforme des mots en **vecteurs numériques** de taille fixe) et d'encoder les mots sous forme d'entiers.

Nous devons également connaître la longueur maximale des séquences d'entrée comme entrée pour le modèle et compléter toutes les séquences à la longueur fixée.

Nous devons également connaître la taille du vocabulaire pour la couche d'embedding.

Enfin, nous pouvons encoder en nombre entier et compléter le texte propre de la description de chaque article.

Définition du modèle

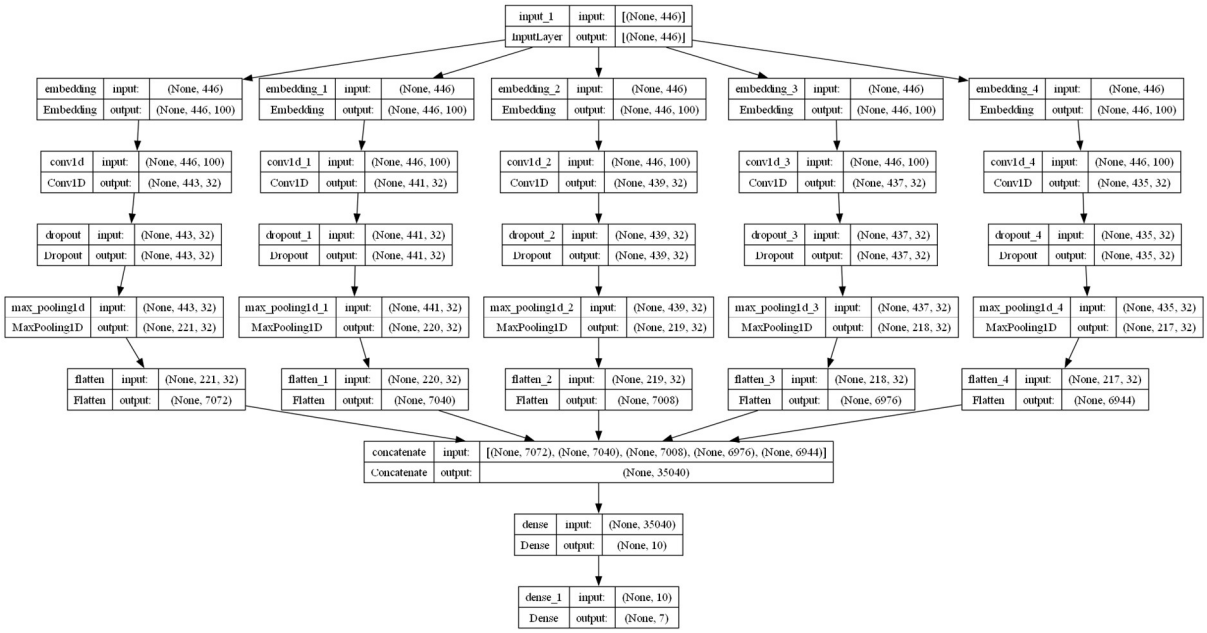
Nous définirons un modèle avec cinq canaux d'entrées pour le traitement de 4, 6, 8, 10 et 12 grammes de texte de description des articles.

Chaque canal est composé des éléments suivants :

- Couche d'entrée qui définit la longueur des séquences d'entrée.
- Couche d'intégration définie sur la taille du vocabulaire et des représentations à valeurs réelles à 100 dimensions.
- Couche convolutive unidimensionnelle avec 32 filtres et une taille de noyau définie en fonction du nombre de mots à lire à la fois.
- Couche de pooling maximale pour consolider la sortie de la couche convolutive.
- Aplatir la couche pour réduire la sortie tridimensionnelle à deux dimensions pour la concaténation.

La sortie des cinq canaux est concaténée en un seul vecteur et traitée par une couche dense et une couche de sortie.

Voici un visuel du modèle



Métrique d'évaluation

Les métriques retenues seront le score F1, la précision ainsi que la durée d'apprentissage.

Ils sont utilisés dans des problèmes de classifications multi classes :

- La précision est utilisée lorsque les vrais positifs et les vrais négatifs sont plus importants tandis que le score F1 est utilisé lorsque les faux négatifs et les faux positifs sont cruciaux
- La précision peut être utilisée lorsque la distribution des classes est similaire tandis que le score F1 est une meilleure mesure lorsqu'il existe des classes déséquilibrées comme dans la plupart des cas.

- Dans la plupart des problèmes de classification réels, une distribution de classe déséquilibrée existe et le score F1 est donc une meilleure mesure pour évaluer notre modèle.

Synthèse des résultats

Nous utiliserons comme modèle de référence Naïve Bayes.

Évaluation

Les deux modèles testés, **Naïve Bayes** et CNN **Multichannel**, ont montré de bonnes performances pour classer les produits en différentes catégories. Toutefois, le modèle CNN **Multichannel** semble légèrement plus précis, avec un score de **93,33%** contre **92,38%** pour Naïve Bayes. Bien que cette différence soit faible, elle indique que le modèle CNN a une meilleure capacité à faire des prédictions correctes.

En regardant plus en détail, on remarque que **CNN fonctionne mieux sur certaines catégories**, notamment "Baby Care" et "Kitchen & Dining", où il obtient un meilleur score de performance. En revanche, **Naïve Bayes est légèrement meilleur** pour la catégorie "Home Decor & Festive Needs". Pour les autres catégories, les deux modèles obtiennent des résultats similaires, ce qui signifie qu'ils sont globalement efficaces pour classer ces types de produits.

L'un des avantages du modèle **Naïve Bayes** est qu'il est plus **simple et rapide** à entraîner. Cela peut être un atout si l'on recherche une solution efficace sans nécessiter trop de puissance informatique. En revanche, le modèle CNN **Multichannel** est plus avancé et capable de repérer des motifs complexes, ce qui peut lui permettre d'obtenir une meilleure précision, même si son entraînement prend plus de temps.

Si l'on devait choisir entre ces deux modèles, **CNN serait un meilleur choix pour maximiser la précision**, notamment si l'on dispose de suffisamment de ressources informatiques. Cependant, si l'objectif est d'avoir un modèle rapide et facile à mettre en place, Naïve Bayes reste une très bonne alternative avec des résultats presque aussi performants.

Les limites et les améliorations possibles

Pour aller plus loin, une analyse plus détaillée des erreurs commises pourrait être utile. Par exemple, grâce à la **matrice de confusion** , on pourrait voir quelles catégories sont les plus souvent confondues et ajuster les modèles en conséquence. Cela permettrait d'améliorer encore la qualité des prédictions et d'optimiser le choix du modèle en fonction des besoins spécifiques.

Les limitations et les améliorations du modèle CNN Multichannel :

Sensibilité aux données d'entrainement . Beaucoup de données sont nécessaires pour bien généraliser. Si le dataset est trop petit ou déséquilibré, le modèle peut apprendre des biais et mal classifier certaines catégories.

Une optimisation serait possible en utilisant des techniques d'augmentation de données (synonymes, paraphrases, back-translation) et s'assurer d'avoir un dataset bien équilibré.

Un risque de surapprentissage , il peut apprendre **trop bien** les données d'entraînement et avoir du mal à s'adapter aux nouvelles données. Cela réduit sa capacité de généralisation.

Il faudrait ajouter des ***couches de régularisation*** *comme le* ***Dropout****, utiliser une bonne* ***stratégie de validation croisée*** *, et limiter la complexité du modèle.*

Un cout Computationnel élevé , **gourmands** en ressources avec plusieurs canaux d'analyse. Entraîner un Multichannel CNN peut être long et nécessiter un **GPU** puissant.

Nous pourrions optimiser l'architecture avec des ***couches 1D plus légères*** *, réduire la taille des embeddings et utiliser du* ***transfer learning*** *avec des modèles pré-entraînés.*

Références.

Approche de référence:
https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

How to Develop a Multichannel CNN Model for Text Classification

https://machinelearningmastery.com/develop-n-gram-multichannel-convolutional-neural-network-sentiment-analysis/

Convolutional Neural Networks for Sentence Classification

https://arxiv.org/abs/1408.5882