## Méthodologie :

**1/Prétraitement des données :**

- lire un fichier CSV contenant des résumés et des textes d'actualité (from kaggle 'news_summary.csv'
- renommer les colonnes pour la préparation des données T5.
- ajouter le préfixe "summarize:" aux textes sources pour la sommation T5.
- diviser les données en ensembles d'entraînement et de test.

**2/Formation du modèle :**

- utiliser la bibliothèque SimpleT5 pour former un modèle T5 sur les données d'entraînement fournies.
- spécifier les longueurs de jetons, la taille des lots, les époques et l'utilisation du GPU.

**3/Inférence du modèle :**

- charger le modèle entraîné.
- traiter un nouveau jeu de données (data_label_1 et data_label_2).
- ajouter le préfixe "summarize:" au texte propre.
- utiliser le modèle pour générer des résumés à partir du texte propre.

**4/Extraction de mots-clés avec TextRank :**

- implémenter l'algorithme TextRank pour extraire des mots-clés des résumés générés.
- La fonction textrank a tokenisé les phrases, supprimé les mots vides et la ponctuation, lemmatisé les mots et construit un graphe basé sur la co-occurrence.
- Les scores PageRank ont été calculés pour chaque mot, et les principaux mots-clés ont été extraits.

**5/Calcul des scores :**

- Une fonction de calcul des scores calculate_final_score a été définie.

Pour chaque résumé généré, la fonction a attribué un score en fonction de la présence de mots-clés pertinents.

NB : Si au moins un quart des mots-clés extraits figuraient dans la liste prédéfinie, le résumé était considéré comme pertinent.

**6/Étiquetage et évaluation :**

- L'étiquette de pertinence pour chaque résumé généré a été déterminée en fonction de la fonction de calcul des scores.
- **Une matrice de confusion** a été générée pour visualiser les performances de la méthode de calcul des scores.

## pourquoi j'ai utilisé ce pretrained TRANSFORMER model T5: text vers text :

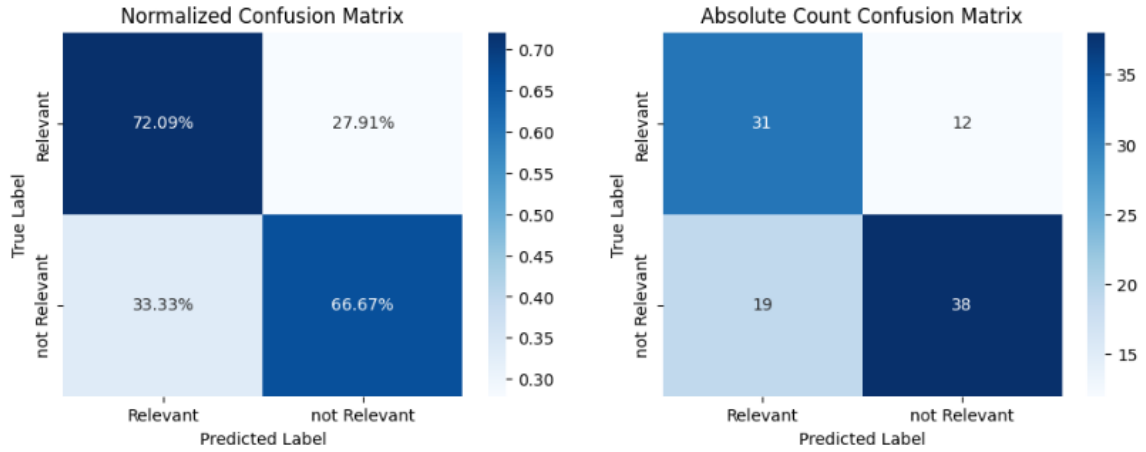
* **Polyvalence** : Le modèle T5 peut être utilisé pour une variété de tâches de traitement du langage naturel, telles que la traduction, la classification, la question-réponse et la génération de résumés. En reformulant chaque tâche en tant que tâche de génération de texte, T5 peut être adapté pour effectuer diverses tâches sans avoir à réentraîner le modèle de manière spécifique.

* **Génération de résumés** : Vous avez utilisé le modèle T5 pour générer des résumés à partir de textes dans votre script. Cette capacité est très utile pour résumer automatiquement de longs documents en versions plus courtes et compréhensibles.

* **Modèle pré-entraîné** : Le modèle T5 est pré-entraîné sur une grande quantité de données textuelles provenant d'Internet, ce qui lui permet d'apprendre les caractéristiques linguistiques et les relations sémantiques. Vous pouvez utiliser cet apprentissage pour des tâches spécifiques sans commencer par zéro en utilisant un modèle pré-entraîné.

* **Transfert d'apprentissage** : Vous pouvez exploiter le transfert d'apprentissage en utilisant un modèle pré-entraîné comme T5. Le modèle comprend déjà le langage naturel et peut être personnalisé en fonction des données de votre tâche, ce qui peut améliorer les performances et réduire les besoins en données d'entraînement.

## RESULTATS :




**Accuracy** : `0.69`

**F-1 score :** `0.6666666666666666`