Rapport sur le Prétraitement des Données

1. Introduction

L'analyse de sentiment vise à déterminer la polarité (positive, négative, neutre) des textes. Pour garantir des résultats précis, il est essentiel d'appliquer des techniques de prétraitement adaptées. Ce rapport présente les méthodes utilisées pour prétraiter le dataset, ainsi que les justifications des choix effectués.

---

2. Analyse Préliminaire du Dataset

Avant toute transformation, le dataset a été analysé pour identifier les problèmes suivants :

- **Valeurs manquantes** : Vérification des champs incomplets.
- **Doublons** : Suppression des entrées répétées pour éviter les biais.
- **Incohérences linguistiques** : Présence de plusieurs langues ou erreurs de saisie.

Bibliothèques utilisées :

- `pandas` : Chargement et exploration des données.

---

3. Techniques de Prétraitement Utilisées

a) Nettoyage du Texte :

- Suppression des caractères répétitifs : réduit les exagérations inutiles (ex : "cooool" → "cool").
- Suppression des caractères accentués : harmonise les termes (ex : "éducation" → "education").
- Suppression des URL, balises HTML et e-mails : élimine les éléments non pertinents pour l'analyse de sentiment.

Bibliothèques utilisées :

- `re` : Manipulation des expressions régulières.
- `nltk` : Suppression des caractères répétitifs.
- `unicodedata` : Suppression des caractères accentués.

b) Normalisation :

- Conversion en minuscules : unifie les mots ("Happy" = "happy").
- Suppression de la ponctuation : réduit le bruit textuel.
- Suppression des stopwords : retire les mots fréquents sans valeur sémantique ("the", "is").

Bibliothèques utilisées :

- `nltk` : Gestion des stopwords.
- `re` : Suppression de la ponctuation.

c) Tokenisation et Lemmatisation :

- **Tokenisation** : segmentation des phrases en mots individuels et en sous-mots (*Subword Tokenization*).
- **Lemmatisation** : réduction des mots à leur forme de base ("running" → "run") pour une meilleure cohérence.

Bibliothèques utilisées :

- `nltk` : Lemmatisation.
- `transformers` (Hugging Face) : Tokenisation en sous-mots.

d) Reconnaissance d'Entités Nommées (NER) :

- Extraction des noms propres, lieux, organisations, etc.
- Permet d'améliorer l'analyse en identifiant les sujets clés du texte.

Bibliothèques utilisées :

- `spaCy` : Reconnaissance d'entités nommées.

---

4. Outils et Bibliothèques Utilisés

Les techniques de prétraitement ont été mises en œuvre à l'aide des outils et bibliothèques suivants :

- **Pandas** : Manipulation et nettoyage des données tabulaires.
- **NumPy** : Gestion des structures de données et calculs numériques.
- **NLTK** : Suppression des stopwords, tokenisation et lemmatisation.
- **spaCy** : Tokenisation avancée et reconnaissance d'entités nommées.
- **re (Regular Expressions)** : Suppression des URL, e-mails et caractères spéciaux.
- **transformers (Hugging Face)** : Tokenisation en sous-mots (*Subword Tokenization*) avec des modèles pré-entraînés comme BERT.
- **unicodedata** : Normalisation des caractères accentués.

---

5. Justification des Choix de Prétraitement

Technique	Objectif	Impact
Suppression des doublons	Éviter les biais de fréquence	Précision accrue de l'analyse
Lemmatisation	Normalisation des mots	Amélioration de la cohérence sémantique
Suppression des stopwords	Concentration sur les mots significatifs	Réduction du bruit dans les données
Conversion en minuscules	Uniformisation	Réduction des doublons inutiles
Suppression de la ponctuation	Nettoyage	Simplification du texte pour l'analyse
Tokenisation en sous-mots	Améliorer la représentation des mots	Meilleure prise en charge des néologismes
Reconnaissance d'entités nommées	Identifier les éléments clés du texte	Amélioration du contexte d'analyse

---

6. Conclusion

Le prétraitement des données est une étape cruciale pour garantir la fiabilité des résultats d'analyse de sentiment. Les méthodes adoptées dans ce rapport ont permis de nettoyer, normaliser et optimiser le dataset, préparant ainsi le terrain pour des analyses précises et significatives. L'ajout de techniques avancées comme la tokenisation en sous-mots et la reconnaissance d'entités nommées améliore la qualité des représentations textuelles et la pertinence des modèles d'analyse de sentiment.